STRUCTURAL AND FUNCTIONAL ADVANCES IN THE EVOLUTIONARY STUDIES OF CELLS AND VIRUSES

BY

ARSHAN NASIR

DISSERTATION

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Informatics
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2015

Urbana, Illinois

Doctoral Committee:

Professor Gustavo Caetano-Anollés, Chair
Associate Professor, Emeritus Jay Mittenthal
Associate Professor Matthew Hudson
Assistant Professor Jian Ma

ABSTRACT

Phylogenomics aims to describe evolutionary relatedness between organisms by analyzing genomic data. The common practice is to produce phylogenomic trees from molecular information in the sequence, order and content of genes in genomes. These phylogenies describe the evolution of life and have become valuable tools for taxonomy. The recent availability of structural and functional data for hundreds of genomes now offer the opportunity to study evolution using more conserved sets of molecular features. Here we report a phylogenomic (i.e. historical) and comparative (ahistorical) analysis that yields novel insights into the origin of cells (Chapters 1-3) and viruses (Chapters 4-6). We utilized conserved protein domain structure information (fold families [FFs] and fold superfamilies [FSFs]) and ontological definitions of gene products (Gene Ontology [GO]) to reconstruct rooted trees of life (ToL), taking advantage of a genomic census of molecular structure and function in the genomes of sampled organisms and viruses. The analysis revealed a global tendency in the proteomic repertories of cellular organisms to increase domain abundance. ToLs built directly from the census of molecular functions confirmed an early origin of Archaea relative to Bacteria and Eukarya, a conclusion further supported by comparative analysis. The analysis further revealed an ancient history of viruses and their evolution by gene loss. Despite the very high levels of variability seen in the replication strategies, morphologies, and host preferences of extant viruses, we recovered a conserved and ancient structural core of protein domains that was shared between cellular organisms and distantly related viruses. This core together with an analysis of the evolution of virion morphotypes strongly suggests an ancient origin for the viral supergroup. Moreover, a large number of viral proteins lacked cellular homologs and strongly negated the idea that viruses merely evolve by acquiring cellular genes. These virus-specific proteins confer pathogenic abilities to viruses and appeared late in evolution suggesting that the shift to parasitic mode of life happened later in viral evolution. The strong evolutionary association between viruses and cells is likely reminiscent of their ancient co-existence inside primordial cells. Moreover, the crucial dependency of viruses to replicate in an intracellular environment creates fertile grounds for genetic innovation. Interestingly, protein domains shared with viruses were widespread in the proteomes of all three cellular superkingdoms suggesting that viruses mediate gene transfer and crucially enhance biodiversity. The phylogenomic trees identify viruses as a ‘fourth supergroup’ along with cellular superkingdoms, Archaea, Bacteria, and Eukarya. The

new model for the origin and evolution of viruses and cells is backed by strong molecular data and is compatible with the existing models of viral evolution. Our experiments indicate that structure and functionomic data represent a useful addition to the set of molecular characters used for tree reconstruction and that ToLs carry in deep branches considerable predictive power to explain the evolution of living organisms and viruses.

ACKNOWLEDGEMENTS

I am extremely grateful to Professor Dr. Gustavo Caetano-Anollés for his kind support and guidance throughout my degree. This dissertation would not have been possible without his enthusiastic support and vast knowledge that he always generously shared with me. Thank you for editing numerous drafts of my dissertation and research papers. Thanks also for letting me explore different research avenues, for backing my ideas, and for helping me grow as a scientist. I simply could not have wished for a better advisor. I would also like to thank my committee members Drs. Jay Mittenthal, Matthew Hudson, and Jian Ma for their constructive criticism, hard questions, and significant input that helped shaped this dissertation. My sincere thanks go to Dr. Kyung Mo Kim for his good advice and friendship that has been beneficial both on a personal and academic level and to Ms. Karin Readel and Dr. Guy Garnett for creating a very productive learning environment at the Illinois Informatics Institute. Words cannot describe the level of support I received from my family, my mother Shaheen Nasir, my late father Muhammad Nasir Qureshi, my sister Jawaria Nasir, and my younger brother Zohaib Nasir. I am also grateful to a number of very special friends, Afraz Ahmed Raja, Masood Kayani, Waqar Ahmed, Yasir Satti, Amir Ansari, Zeeshan Fazal, Malik Nadeem Akhtar, Syed Abbas Bukhari, Ahmed Sadeque, Tayyab Nawaz, Gary Umphrey, Aaron Bird, Abraham Akpertey, Sumin Kim, Khuram Shahzad, Iftikhar Ahmed, Shahzad Bhatti, Fayez Aziz, Waseem Haider, Fizza Mughal, Momal Arslan, Farah Zaib Khan, Rafia Khan, Sana Zeeshan, Sonya Gulzeb, Adnan Ansari, Umair Seemab, Faran Haq Jahangiri, Megan Johns, Salman Salahuddin Ayubi, Daniel Wong, Salman Khan, Jamal Hussain, and Usman Saeed for being part of this journey. Special thanks for the generous funding support that came from the COMSATS Institute of Information Technology, National Science Foundation, Illinois Informatics Institute, Chateaubriand Association, and the Dissertation Completion Fellowship from the Graduate College of the University of Illinois. In the end, I consider myself fortunate to have worked with the most amazing group of scientists at the Illinois Informatics Institute and *Evolutionary Bioinformatics Laboratory.*

TABLE OF CONTENTS

PREFACE

Chapter 1 describes the evolutionary dynamics of protein domain gain and loss in cellular organisms and how this affects their proteomic make up and long-term evolution. We retraced the history of changes in the abundance and occurrence of FF domains along the many branches of the ToL and inferred global patterns of protein domain gain and loss. Results revealed that both gains and losses were frequent events in the evolution of cells. However, gains generally overshadowed the number of losses. This trend was consistent in the three superkingdoms, albeit at unequal rates. Interestingly, the gain-to-loss ratios were much higher in akaryotes (prokaryotes) compared to eukaryotes suggesting ongoing secondary adaptions in their evolution. Functional annotations of FF domains revealed that both Archaea and Bacteria gained and lost metabolic capabilities during the course of evolution while Eukarya acquired a number of diverse molecular functions including those involved in extracellular processes, immunological mechanisms, and cell regulation. The increasing number of domain gains in proteomes is predicted to redefine the persistence strategies of organisms in superkingdoms, influence the make up of molecular functions, and enhance organismal complexity by the generation of new domain architectures.

Chapter 2 presents novel phylogenomic trees and networks of organisms inferred directly from the GO annotations. Phylogenies and networks yielded significant insights into the emergence and evolution of cellular life. Based on this analysis we conclude that the ancestor of Archaea originated earlier than the ancestors of Bacteria and Eukarya and was thermophilic. In contrast, basal bacterial lineages were non-thermophilic. A close relationship between Plants and Metazoa was also identified that disagrees with the traditional Fungi-Metazoa and Plants-Fungi groupings. While measures of evolutionary reticulation were minimum in Eukarya and maximum in Bacteria, the massive role of horizontal gene transfer in microbes did not bias phylogenomic reconstructions. Phylogenies and networks also revealed that the best reconstructions were recovered when problematic taxa (i.e., parasitic/symbiotic organisms) and horizontally transferred characters were excluded from the analysis.

The historical analyses of Chapters 1 and 2 were supported by another analysis using an ahistorical and comparative proteomic and functionomic inferential framework for genome evolution (Chapter 3) that successfully resolved the tripartite division of cells and sketched their

history, without employing any phylogenetic algorithm. Here, evolutionary inferences were derived directly from the spread of conserved molecular features, such as protein structures and molecular functions, in the proteomes and functionomes of contemporary organisms. Patterns of use and reuse of these traits uncovered a strong evolutionary association between Bacteria and Eukarya and revealed marked evolutionary reductive tendencies in the archaeal genomic repertoires. Our study highlights a strong vertical trace in the history of proteins and associated molecular functions, which was reliably recovered using the comparative genomics approach. The trace supported the existence of a stem line of descent and the very early appearance of Archaea as a diversified superkingdom, but failed to uncover a hidden canonical pattern in which Bacteria was the first superkingdom to deploy superkingdom-specific structures and functions.

The comparative and phylogenomic approaches were extended to include viruses into the evolutionary picture (Chapter 4). This exercise uncovered unprecedented and remarkable trends in the evolution of viruses and cells. Viral proteomes harbored a large number of FSF domains that lacked cellular homologs. Moreover, they shared a variety of metabolic and informational FSFs with cellular organisms. The ancient history and co-existence of viruses with ancient cells was confirmed by both the comparative and phylogenomic analysis. The analysis revealed that modern viruses originated from ancient cells that harbored segmented 'viral-like' RNA genomes. These ancient cells eventually reduced into modern day viruses while their siblings diversified into Archaea, Bacteria, and Eukarya. The redefined ToL identifies viruses as a distinct and ancient supergroup that played important roles during the evolution of cells. We also propose that the crucial dependency of viruses to replicate inside cellular hosts creates a rich environment for evolutionary innovation and likely benefits the long-term evolution of modern cells. This was demonstrated by the distribution of viral replicon types in host organisms and by the physiological and molecular make up of modern cellular organisms (Chapter 5). Finally, Chapter 6 briefly reviews current research on viral evolution and gives direction for future analysis in hope to benefit our understanding of viruses and their impact on the evolution of cells.

The dissertation is divided into six main chapters. Chapters 1-4 are organized as typical research articles and start with an introductory section that defines the rationale of study and provides background information on the topic. While chapters 5 and 6 focus specifically on viral origins and suggest promising directions for further research in this area.

CHAPTER 1: GLOBAL PATTERNS OF PROTEIN DOMAIN GAIN AND LOSS IN SUPERKINGDOMS[1]

Introduction

Proteins are biologically active molecules that perform a wide variety of functions in cells. They are involved in catalytic activities (e.g. enzymes), cell-to-cell signaling (hormones), immune response initiation against invading pathogens (antibodies), decoding genetic information (transcription and translation machinery), and many other vital cellular processes (receptors, transporters, transcription factors). Proteins carry out these functions with the help of well-packed structural units referred to as domains [1]. Domains are modules within proteins that can fold and function independently and are evolutionarily conserved [2-4]. It is the domain make up of the cell that defines its molecular activities and leads to interesting evolutionary dynamics [5].

Different mechanisms have been described to explain the evolution of domain repertoires in cells [3]. These include the reuse of existing domains [2,6], interplay between gains and losses [7-9], *de novo* domain generation [1], and horizontal gene transfer (HGT) [10]. Domains that appeared early in evolution are generally more abundant than recently emerged domains and can be reused in different combinations in proteins. This recruitment of ancient domains is an ongoing evolutionary process that leads to the generation of novel domain architectures (i.e. ordering of domains in proteins) by gene fusion, exon recombination and retrotransposition [11]. For example, aminoacyl-tRNA synthetases are enzymes that charge tRNAs with 'correct' amino acids during translation [12,13]. These crucial enzymes are multidomain proteins that encode a catalytic domain, an anticodon-binding domain, and in some cases, accessory domains involved in RNA binding and editing [13]. Evolutionary analysis suggests that these domains were recruited gradually over time [14]. In fact, recruitment of ancient domains to perform new functions is a recurrent phenomenon in metabolism [15].

In addition to the frequent reuse of domains, the dynamics between gains and losses also impacts the evolution of proteome repertoires [7-9]. Previous studies identified high rates of gene

[1]This chapter has been published as manuscript in *PLoS Computational Biology* (see [317]). The final publication is

gains and losses (even) in 12 very closely related strains of Drosophila [7], Prochlorococcus (a genus of cyanobacteria) [16] and 60 isolates of Burkholderia (a genus of proteobacteria) [17]. A recent analysis of Pfam domains [18] revealed that ~3% of the domain sequences were unique to primates and had emerged quite recently [1,19]. This implies that emergence of novel domains is an incessant evolutionary process.

In contrast, different selective pressures can lead to loss of domains in certain lineages and trigger major evolutionary transitions. For example, the increased rate of domain loss has been linked to reductive evolution of the proteomes of the archaeal superkingdom [20], adaptation to parasitism in cells [21] (e.g. transition from the free-living lifestyle to obligate parasitism in *Rickettsia* [22]), and 'de-evolution' of animals [23,24] from their common ancestor. In these studies, gain and loss dynamics were inferred for only particular groups of phyla or organisms. A global analysis involving proteomes from the three superkingdoms remained a challenge. Finally, changes to domain repertoires are also possible by HGT that is believed to occur with high frequency in microbial species, especially Bacteria [25,26].

Here, we describe the evolutionary dynamics of protein domains grouped into fold families (FFs) and model the effects of domain gain and loss in the proteomes of 420 free-living organisms that have been fully sequenced and were carefully sampled from Archaea, Bacteria, and Eukarya. The 420-proteome dataset was previously used by our group to reconstruct the evolutionary history of free-living organisms (see [27]) and was updated here to account for recent changes in protein classification and functional annotation. The dataset is very well annotated, especially regarding organism lifestyles that are otherwise problematic to assign, has already produced patterns of protein and proteome evolution that are very useful (including those described in [27]), and has produced timelines of FF evolution that are being actively mined.

We conducted phylogenomic analyses using the *abundance* (i.e. total redundant number of each FF in every proteome) [28,29] and *occurrence* (presence or absence) [30,31] counts of FFs as phylogenetic characters to distinguish the 420 sampled taxa (proteomes). FF information was retrieved from the Structural Classification of Proteins (SCOP) database, which is considered a 'gold standard' for the classification of protein domains into different hierarchical levels [32]. Current SCOP definitions group protein domains with high pair-wise sequence identity (>30%) into a common FF, FFs that are evolutionarily related into fold superfamilies

(FSFs), FSFs with similar secondary structure arrangement into folds (Fs), and Fs with common secondary structure elements into a handful of protein classes [33,34]. A total of 110,800 SCOP domains (ver. 1.75) are classified into a finite set of only 1,195 Fs, 1,962 FSFs and 3,902 FFs. The lower number of distinct FSFs and FFs suggests that domain structure is far more conserved than molecular sequence (e.g. see [35]) and is reliable for phylogenetic studies involving the systematic comparison of proteomes [27]. Another advantage of using SCOP domains is the consideration of known structural and inferred evolutionary relationships in classifying domains into FFs and FSFs [36]. In comparison, evolutionary relationships for the majority of the Pfam domains are unknown.

We further restricted the analysis to include only FF domains as they are conserved enough to explore both the very deep and derived branches of the tree of life (ToL) and are functionally orthologous [37]. In contrast, FSF domains represent a higher level in SCOP hierarchy and are more conserved than FFs but may or may not be functionally orthologous. Moreover, high conservation of FSF domains is useful for exploring the deep branches of the ToL but may not be very informative for the more derived relationships.

The analysis of retracing the history of changes in the occurrence and abundance of FF domains on each branch of the reconstructed ToLs revealed that FFs were subject to high rates of gains and losses. Domain gains generally outnumbered losses but both occurred with high frequencies throughout the evolutionary timeline and in all superkingdoms. Remarkably, the gains-to-loss ratios increased with evolutionary time and were relatively higher in the late evolutionary periods. Finally, functional annotations of FFs illustrated significant differences between superkingdoms and described modern tendencies in proteomes.

Methods

Data retrieval and processing

The 420-proteome dataset used in this study included proteomes from 48 Archaea, 239 Bacteria, and 133 Eukarya. The dataset did not include any parasitic organisms as they harbor reduced proteomes and bias the global phylogenomic analyses [21,38]. FFs were assigned to proteomes using SUPERFAMILY ver. 1.73 [39,40] hidden Markov models (HMMs) [41] at an *E*-value cutoff of 10^{-4} [42]. A total of 2,397 significant FF domains were detected in the sampled proteomes. The definitions of 8 FFs in the 420-proteome dataset were updated in SCOP ver. 1.75 and were therefore renamed in our dataset. FFs were referenced using SCOP *concise classification strings (css)* (e.g. 'Ferredoxin reductase FAD-binding domain-like' FF is b.43.4.2, where b represents the class [all-beta proteins], 43 the fold, 4 the FSF and 2 the FF).

Phylogenomic analysis

We considered the genomic *abundance* [28,29] and *occurrence* [30,31] of 2,397 FFs as phylogenetic characters to reconstruct phylogenies describing the evolution of 420 free-living organisms (i.e. taxa) using maximum parsimony (MP). The raw abundance values of each FF in every proteome (g_{ab}) were log-transformed and divided by the logarithm of maximum value in the matrix (g_{max}) to account for unequal proteome sizes and variances (see formula below) [29,43].

$$g_{ab\_normal} = \text{round}\,[\ln(g_{ab} + 1) / \ln(g_{max} + 1) * 23]$$

The transformed abundance values were then rescaled from 0 to 23 (scaling constant) in an alphanumeric format (0-9 and A-N) to allow compatibility with the phylogenetic reconstruction software. The transformed abundance matrix with 24 possible character states was imported into PAUP* 4.0b10 [44] for the reconstruction of *abundance* trees. For *occurrence* trees, we simply used 0 and 1 (indicating absence and presence) as the valid character state symbols.

We polarized both *abundance* and *occurrence* trees using the ANCSTATES command in PAUP* and designated character state 0 as the ancestral state, since the most ancient proteome is closer to a simple progenote organism that harbors only a handful of domains [20,38]. The stem lineage of this organism gradually increased its domain repertoire, supporting the polarization

from 0 to N and Weston's generality criterion, in which the taxic distribution of a set of character states is a subset of the distribution of another [45,46].

Phylogenetic trees are adequately interpreted when rooted. This provides direction to the flow of evolutionary information and is useful to study species adaptations. In this study, we choose to root trees using the Lundberg method [47]. This scheme first determines the most parsimonious unrooted tree, which is then attached to a hypothetical ancestor. The hypothetical ancestor may be attached to any of the branches in the tree. However, only the branch that gives the minimum increase in overall tree length is selected [48]. This branch, which exhibits the largest numbers of ancestral (plesiomorphic) character states was specified using the ANCSTATES command in PAUP*. Thus, Lundberg rooting automatically roots the trees by preserving the principle of MP. This method is simple and free from artificial biases introduced by alternative rooting methods (e.g. the outgroup method). While selection of an appropriate outgroup to root the ToL is virtually impossible, Lundberg rooting provides a parsimonious estimate of the overall phylogeny and should be considered robust as long as the assumptions used to root the trees are not proven false. To evaluate support for the deep branches of ToLs, we ran bootstrap (BS) analysis with 1,000 replicates. Character state changes were recorded by specifying the 'chglist' option in PAUP*. Trees were visualized using Dendroscope ver. 3.0.14b [49].

Tree comparison

To determine congruence between *abundance* and *occurrence* trees, we used the nodal module implemented in the TOPD/FMTS package (ver. 3.3) [50]. The module takes as input a set of trees in Newick format and calculates a root mean squared deviation (RMSD) value for each pairwise comparison. The RMSD value is 0 for identical trees and increases with incongruence. To evaluate the significance of calculated RMSD values, we implemented the 'Guided randomization test' with 100 replications to determine whether the calculated RMSD value was smaller than the chance expectation. The randomization test randomly changes the positions of taxa in trees, while maintaining original tree topology, and calculates an RMSD value for each random comparison [50]. The result is a random distribution of RMSD values with a mean and standard deviation. The calculated RMSD value was compared with the mean

of the random distribution to determine whether the observed differences were better than what would be expected merely by chance.

Spread (popularity) of FFs in proteomes

The spread of each FF was given by its distribution index (*f*-value), defined by the total number of proteomes encoding a particular FF divided by the total number of proteomes. The *f*-value ranges from 0 (absence from all proteomes) to 1 (complete presence).

Molecular and geological age of FFs

To determine the relative age of FF domains in our dataset, we reconstructed trees of domains (ToDs) from the *abundance* and *occurrence* matrices used in the reconstruction of ToLs. The matrices were transposed, treating FFs as taxa and proteomes as characters. The reconstructed ToDs described the evolution of domains grouped into FFs and identified the most ancient and derived FFs (refer to [27] for an elaborate description and discussion on ToDs). To root the trees, we declared character state 'N' as the most ancestral state. This axiom of polarization considers that history of change for the most part obeys the 'principle of spatiotemporal continuity' (*sensu* Leibnitz) that supports the existence of Darwinian evolution. Specifically, it considers that abundance and diversity of individual FFs increases progressively in nature by gene duplication (and associated processes of subfunctionalization and neofunctionalization) and *de novo* gene creation, even in the presence of loss, lateral transfer or evolutionary constraints in individual lineages. Consequently, ancient domains have more time to accumulate and increase their abundance in proteomes. In comparison, domains originating recently are less popular and are specific to fewer lineages. We note that the N to 0 polarization is supported by the observation that FFs that appear at the base of the ToDs are structures that are widespread in metabolism and are considered to be of very ancient origin (e.g. [27]).

The age of each FF was drawn directly from the ToDs using a PERL script that calculates the distance of each node from the root. This node distance (*nd*) is given on a relative scale and portrays the origin of FFs from 0 (most ancient) to 1 (most recent). The geological ages of FFs were derived from a molecular clock of protein folds [51,52] that was used to calibrate important events in proteome evolution. We have previously shown that *nd* correlates with geological time, following a molecular clock that can be used as a reliable approximation to date the appearance of protein domains [51,52].

Functional annotations

We used the SUPERFAMILY functional annotation scheme (based on SCOP 1.73) to study the functional roles of FF domains in our dataset [53-55]. The SUPERFAMILY annotation assigns a single molecular function to FSF domains (and by extension to its descendant FFs). The annotation scheme gives a simplified view of the functional repertoire of proteomes using seven major functional categories including, i) *metabolism*, ii) *information*, iii) *intracellular processes*, iv) *extracellular processes*, v) *general*, vi) *regulation* and vii) *other* (includes domains with either unknown or viral functions). We assumed that FFs grouped into an FSF performed the same function that was assigned to their parent FSF [27]. While this simplistic representation does not demonstrate the complete functional capabilities of a cell, it is sufficient to illustrate the major functional preferences in proteomes (refer to [21] for further description and use of the functional annotation scheme in large-scale proteomic studies).

Gene Ontology (GO) enrichment analysis

We conducted a GO enrichment analysis [56,57] on FF domains to identify biological processes that were significantly enriched. For this purpose, the list of FF domains was given as input to domain-centric Gene Ontology (dcGO; http://supfam.org/SUPERFAMILY/dcGO) resource and the most specific and significant associations to GO terms corresponding to different biological processes [58,59] were retrieved. The statistical significance was evaluated by *P*-value computed under the hypergeometric distribution [56], while the false discovery (*FDR*) rate was set to default at $< 10^{-2}$ [60].

Results

We first describe the patterns of FF use and reuse in superkingdoms and then build on this knowledge to infer the meanings of domain gain and loss in proteomes.

Evolutionary history of FF domains

A Venn diagram describes the sharing patterns of 2,397 FFs in seven Venn distribution groups (Figure 1.1A). For simplicity, we name these sets 'taxonomic groups' with the understanding that their taxonomic status is endowed by patterns of distribution of FFs in superkingdoms. The number of FFs decreased in the order Eukarya (total FFs = 1,696), Bacteria (1,510) and Archaea (703). Eukarya also had the highest number of superkingdom specific FFs (758), followed by Bacteria (522), and Archaea (89). ABE FFs were universal (i.e. present in all three superkingdoms) and made the third largest group with 484 FFs, while BE was the fourth largest taxonomic group with 414 FFs (Figure 1.1A). The lowest number of FFs was in AE with only 40 FFs that were unique to both Archaea and Eukarya. The number of Archaea-specific FFs was also low (89) but comparable to the number of akaryotic (prokaryotic) FFs (i.e. AB = 90). We observed that Archaea was mostly about sharing (or not innovating new FFs). This was evident by the fact that only 13% of the total archaeal FFs were Archaea-specific. This was in striking contrast with Bacteria and Eukarya where superkingdom-specific FFs made large proportions of the FF repertoires with 35% and 45% FFs, respectively (Figure 1.1A).

We plotted the distribution of domain ages (i.e. *nd*) for FFs in each taxonomic group to determine the order of their evolutionary appearance (Figure 1.1B) (see Methods). The first FF to appear in evolution was the 'ABC transporter ATPase domain-like' (c.37.1.12) FF at *nd* = 0 in the ABE taxonomic group (Figure 1.1B). ABC transporters are multifunctional proteins that are primarily involved in the transport of various substrates across membranes [61,62]. These domains are ubiquitous and highly abundant in extant species and considered to be very ancient. In our timeline, c.37.1.12 appeared first, supporting its widespread presence and significance in cells. ABE was the most ancient taxonomic group spanning the entire time axis with a median *nd* of 0.24 (Figure 1.1B). This suggested that the majority of the FFs that were common across all superkingdoms appeared very early in evolution.

ABE was followed by the appearances of BE (at *nd* = 0.15), AB (0.26), B (0.26), E (0.551), A (0.555), and AE (0.57) taxonomic groups, in that order (Figure 1.1B). The first

complete loss event for any FF in the primordial world likely triggered the appearance of the BE taxonomic group. Our data indicates that this occurred at *nd* = 0.15 (roughly >3.2 billion [Gyrs] years ago] with the complete loss of the 'Heat shock protein 90, HSP90, N-terminal domain' (d.122.1.1) FF in Archaea (Figure 1.1B). Heat-shock proteins are molecular chaperones that assist in protein folding and clearing of cell debris [63]. These are highly conserved in bacterial and eukaryal species, but relatively less abundant in Archaea. In fact, homologs of Hsp90 or Hsp100 are completely absent in archaeal species [63]. This knowledge is compatible with our finding of loss of d.122.1.1 FF in Archaea that occurred very early in evolution. We propose that this event exemplifies reductive evolutionary processes that were at play early in evolution in nascent archaeal lineages as emergent diversified cells were unfolding different mechanisms of protein folding. In light of our results, Archaea was the first superkingdom to follow reductive trends (read below).

The first superkingdom-specific FF appeared in B at *nd* = 0.26 (~2.8 Gyrs ago), while both Archaea and Eukarya acquired unique FF domains concurrently at around *nd* = 0.55 (~1.6 Gyrs ago) (Figure 1.1B). Emergence of taxonomic groups in evolution described three important evolutionary epochs: (i) *early* ($0 \leq nd < 0.15$), a period before the start of reductive evolution in the archaeal superkingdom, (ii) *intermediate* ($0.15 \leq nd < 0.55$), a period marked by early domain discovery in Bacteria, and (iii) *late* ($0.55 \leq nd \leq 1$), a period during which simultaneous diversification of Archaea and Eukarya occurred (Figure 1.1B).

To determine the popularity (spread) of FFs across organisms, we computed an *f*-value representing the fraction of proteomes encoding an FF. The median *f-v*alue decreased in the order, ABE > AE > E > BE > AB > A > B (Figure 1.1C). We observed that universal FFs of the ABE taxonomic group were most popular and shared by the majority of the proteomes (median *f* = 0.58). The FFs in AE and E were also distributed with higher *f*-values (median *f* = 0.54 and 0.27). In contrast, most of the bacterial taxonomic groups (e.g. BE, AB and B) had lower median *f*-values (0.22, 0.10, and 0.02, respectively). The Venn diagram indicated that 522 out of 2,397 FFs were bacteria-specific (Figure 1.1A) but the median *f*-value of those FFs was extremely low (0.02) (Figure 1.1C). This implies that FFs unique to Bacteria were very unevenly distributed among bacterial species. This also suggested that the rate of FF discovery in Bacteria was very high but their spread was quite limited.

A recent study proposed concepts of economy (i.e. organism budget in terms of number of unique genes and domain structures), flexibility (potential of an organism to adapt to environmental change) and robustness (ability to resist damage and change) to help explain the persistence strategies utilized by organisms in the three superkingdoms [64]. To determine how persistence strategies distributed in our dataset, we redefined economy (total number of unique FFs in a proteome), flexibility (total number of redundant FFs in a proteome) and robustness (ratio of flexibility to economy). When plotted together on a 3D plot, interesting patterns were revealed (Figure 1.1D). As expected, the proteomes of the akaryotic microbes in Archaea and Bacteria were most economical but least flexible and robust (Figure 1.1D). Within these superkingdoms, archaeal proteomes (red circles) exhibited greatest economy but lowest flexibility and robustness. In contrast, Bacteria exhibited intermediate levels of economy, flexibility and robustness. Finally, eukaryal proteomes were least economical but highly flexible and robust (Figure 1.1D). Table 1.1 lists the lower and upper bounds for economy, flexibility, and robustness for the three superkingdoms. The median values for the three parameters always increased in the order, Archaea, Bacteria, and Eukarya (Table 1.1). The analysis revealed that the survival strategy of microbial species lies in encoding smaller domain repertoires while the eukaryal species trade-off economy with more flexibility and robustness and harbor richer proteomes [64]. The number of both unique (economy) and redundant FFs (flexibility and robustness) was considerably higher in eukaryotes.

Functional annotation of FF domains in history

Using the SUPERFAMILY functional annotation scheme [53-55], we were able to compare the distributions of molecular functions in taxonomic groups (Figure 1.2A) and date their evolutionary appearance (*nd*) (Figure 1.2). *Metabolism* was the most abundant and widely distributed molecular function in organisms, especially in the ABE, BE, and AB taxonomic groups. However, significant deviations were observed in the AE and A taxonomic groups, where informational FFs (e.g. those belonging to the replication machinery) outnumbered FFs in other functional categories (Figure 1.2A). These results are consistent with previous knowledge regarding high sharing of informational proteins between Archaea and Eukarya and a common metabolic apparatus between Bacteria and Eukarya. This observation has often led to proposals relating the origin of eukaryotes to a confluence between akaryotic cells (reviewed in [65]; see also [66-69]). However, our data show that the presence of bacterial metabolic enzymes in

Eukarya is better explained by primordial endosymbiotic events leading to mitochondria and plastids in a proto-eukaryote stem cell-line (read below). In comparison, sharing of informational enzymes between Archaea and Eukarya occurred relatively late in evolution and could actually reflect late domain losses in Bacteria. *Intracellular processes* and *general* were distributed similarly while *regulation* and *extracellular processes* appeared to be preferential only in E (Figure 1.2A). The distribution of molecular functions in taxonomic groups was largely in agreement with the distribution previously explained for individual species [21].

We explored the order of evolutionary appearance of molecular functions by generating *nd* vs. *f* plots for the seven taxonomic groups (Figure 1.2). The ABE FFs were present with largest *f*-values and as expected spanned the entire *nd*-axis (Figure 1.2B). In fact, 13 FFs had an *f*-value of 1.0 indicating universal presence in organisms, while 62 near-universal FFs were present in >95% of the proteomes. ABE FFs were generally enriched in metabolic functions (Figure 1.2B). This suggested that the last common ancestor of diversified life was structurally and metabolically versatile (e.g. [38]). However, the *f*-value distribution of ABE FFs followed a bimodal pattern with a significant drop in *f* during the *intermediate* evolutionary epoch. Most of the FFs of intermediate age were classified as metabolic (grey circles), informational (red circles), or with intracellular roles (light blue circles) (Figure 1.2B).

BE followed a distribution similar to ABE but the first FF appeared during the *intermediate* evolutionary epoch at *nd* = 0.15 (Figure 1.2C). This also marked the first loss of an FF in Archaea (boxplot for BE in Figure 1.1B). This observation implies that Archaea was the first superkingdom to escape from the ancestral community and evolved by streamlining genomes. Perhaps, genome reduction was better suited for harsher environments. Other selective pressures that may have triggered early domain loss in Archaea could include escape from RNA viruses (RNA is unstable at extreme temperatures) and phagotrophs [70]. The majority of the BE FFs served metabolic, informational and intracellular roles (Figure 1.2C), just like ABE.

The akaryotic-specific (AB) FFs appeared during the *intermediate* and *late* evolutionary epochs and were largely dominated by metabolic and *other* FFs (Figure 1.2D). Most of these FFs had very low *f*-values (Figure 1.2D) indicating that this taxonomic group exhibited low popularity levels. In contrast, all of the 40 AE FFs appeared in the *late* epoch and were dominated by domains involved in informational (red) (Table 1.2) and regulatory processes

(green) (Figure 1.2E). This validated the hypothesis that informational enzymes in eukaryotes are very similar to their archaeal counterparts rather than bacterial enzymes [71-73]. This argument has been used to propose a sister relationship between Archaea and Eukarya and an ancient origin of Bacteria. However, our analysis revealed that sharing of informational domains between archaeal and eukaryal species was only a recent event (i.e. was evident in the *late* evolutionary epoch; $nd \geq 0.55$) and that the canonical sister relationship between Archaea and Eukarya inferred from the 16S rRNA trees [74] is influenced by the high rates of modern sharing between Archaea and Eukarya (read below) [75]. AE FFs were generally distributed with higher *f*-values (Figure 1.2E).

FFs unique to Archaea (A) appeared in the *late* epoch at $nd = 0.55$ and were generally distributed with lower *f*-values (Figure 1.2F). The discoveries of these FFs were biased towards informational and *other* domains (Figure 1.2F). A large number of bacteria-specific FFs (B) also appeared during the *intermediate* and *late* evolutionary epochs (Figure 1.2G). We note that, in general, bacterial FFs appearing in the *intermediate* epoch were biased towards informational roles, while those that appeared later served metabolic and intracellular roles (Figure 1.2G). Lastly, all of the Eukarya-specific (E) FFs appeared in the *late* epoch (Figure 1.2H), just like Archaea (Figure 1.2F). Eukarya discovered a large number of recent FF domains ($nd \geq 0.55$) that were involved in regulation (green circles) and extracellular processes (blue circles) and were distributed with relatively high *f*-values in the eukaryal proteomes (Figure 1.2H).

Superkingdom-specific FFs appeared in both Archaea and Eukarya at around the same time, and both showed a tendency to become widespread in species (Figure 1.2H). In contrast, the discovery of Bacteria-specific (B) FFs started much earlier but with limited spread (Figure 1.2G). This suggested that while Archaea was the first superkingdom to follow reductive trends, it was Bacteria that diversified first and was capable of unfolding superkingdom-specific domain structures. The primordial stem-line (that was structurally and functionally complex) later evolved into eukaryotes, possibly after engulfment of already diversified microbes (read below). In this regard, we identified a set of mitochondrial FFs, all of which appeared at $nd \geq 0.55$, during and after the rise of the E taxonomic group, including the 'Mitochondrial resolvase ydc2 catalytic domain' (c.55.3.7; $nd = 0.55$) and the 'Mitochondrial cytochrome c oxidase subunit VIIb' (f.23.5.1; $nd = 0.59$) FFs (Table 1.3). Thus, our timelines do not support fusion hypotheses for the origin of eukaryotes linked to a confluence between akaryotes. The fusion scenarios have

been discussed elsewhere [65,70,76-79] and it is beyond the scope of this study to evaluate what model is better. In light of our data that is based on the genomic census of conserved FF domains in hundreds of free-living organisms, we support a phagotrophic and eukaryote-like nature of the host (anticipated in [78,79]) that acquired the primordial alpha-proteobacterium as an endosymbiont, which later became mitochondria and triggered the diversification of eukaryotes (at *nd* = 0.55; roughly ~1.6 billion years ago). A formal test of this hypothesis is warranted and will be explored in a future study. The exercise also revealed that the lower median *f*-values observed earlier (Figure 1.1C) were due to the significant drop in *f* in the *intermediate* evolutionary epoch. We note that the majority of the bacterial FFs (i.e. belonging to the ABE, BE, B and AB taxonomic groups) also appeared during this period and thus affected the overall medians.

Phylogenomic patterns

We generated rooted ToLs from *abundance* (Figure 1.3A) and *occurrence* (Figure 1.3B) counts of 2,397 FF domains in the 420 free-living proteomes using MP as the optimality criterion in PAUP* 4.0b10 [44]. Both reconstructions recovered a previously established tripartite world of cellular organisms [20,27,74,80]. The archaeal superkingdom always formed a paraphyletic group at the base of the ToLs. The deep branches of the ToLs were occupied by thermophilic and hyperthermophilic archaeal species (*Thermofilum pendens* and *Cand.* Korarchaeum) (Figure 1.3). Rooting of ToLs in Archaea is in conflict with canonical trees that are rooted in Bacteria and are reconstructed from ancient paralogous gene couples and 16S rRNA gene sequences [74,81,82]. Instead, the archaeal rooting is supported by a number of previous studies [14,20,27,83-85].

Bacteria and Eukarya formed strong monophyletic clades that were supported by high BS values (≥ 99%) and were separated from Archaea with 53% (Figure 1.3A) and 78% (Figure 1.3B) BS support. Both ToLs had strong phylogenetic signal (g_1 = -0.33 and -0.28). Overall, phylogenomic patterns resembled traditional groupings and supported previous analyses of similar kind [20,27]. Moreover, the dissimilarity between two reconstructions was 5.37, which was smaller than the mean RMSD calculated from 100 random comparisons (Figure 1.3) (see Methods). Because the ToLs were supported with high confidence and resembled previous

analyses [20,27], they made useful tools for the study of domain gain and loss events on the many branches (read below).

Global patterns of domain gains and losses

To quantify the relative contributions of domain gains and losses impacting the evolution of superkingdoms, we retraced the history of character state changes (i.e. changes in the abundance or occurrence of FFs) on each branch of the reconstructed ToLs. For each FF domain, we counted the number of times it was gained and lost in different branches of the phylogenetic tree. Gains were recorded when the abundance/occurrence of a particular FF at a node was higher than the corresponding value at the immediate ancestral node. In turn, losses were incremented when the abundance/occurrence of a particular FF at a node was lower. Because we allowed character changes in both forward and backward directions (i.e. Wagner parsimony), each FF character could be both gained and lost a number of times across the many branches of the ToL. This assumption is reasonable as different lineages of organisms utilize domain repertoires differently. Because abundance counts are expected to be higher in the eukaryotic species (especially in Metazoa) due to increased gene duplication events and a persistence strategy that favors flexibility and robustness (Figure 1.1D) [64], we also considered gains and loss statistics from the *occurrence* trees.

To evaluate the performance of both models, we first compared the number of FFs that were gained (i.e. net sum above zero) and lost (net sum below zero) in both reconstructions. Out of the total 2,397 (2,262 parsimony informative) FF domains in the *abundance* model, 1,955 (86%) were gained, while only 236 (10%) were lost. In contrast, *occurrence* identified 60.1% FFs as gained (1,353/2,249) and 30.5% (686/2,249) as lost. Nearly 96% (1300/1,353) of the *occurrence* gains were also gained in *abundance* while only 26% (178/686) losses were common to both models. We infer that *abundance* included nearly all the gains from *occurrence* and likely overestimated the number of gains (due to gene duplications and domain reuse). In contrast, *occurrence* led to more balanced distributions but likely overestimated losses (read below).

To provide additional support to the gain/loss model, we pruned taxa from the original ToLs leaving only one superkingdom and recalculated character state changes on the pruned trees. This eliminated any biases resulting from the differences in persistence strategies of the

three superkingdoms and yielded four phylogenetic trees, *Total* (taxa = 420, total FF characters = 2,397), *Archaea* (48, 703), *Bacteria* (239, 1,510) and *Eukarya* (133, 1,696). For each of the four trees, we calculated the sum of gain and loss events for all parsimony informative FF characters and represented the values in boxplots (Figure 1.4A). In all distributions, medians were above 0 indicating that the sum of net gains and losses was a non-negative number for both *abundance* (Figure 1.4A:*abundance*) and *occurrence* (Figure 1.4A:*occurrence*) models. The exception was the eukaryal tree pruned from the *occurrence* model, for which the median was exactly zero. The result revealed that while both gains and losses occurred quite frequently, the former was more prevalent in proteome evolution.

The histograms in Figure 1.4B describe the distributions of gain and loss counts for all parsimony informative FF characters in the *Total* dataset. When plotted against evolutionary time (*nd*), results highlighted remarkable patterns in the evolution of domain repertoires. Domain gains outnumbered losses in both *abundance* (80,904 gains vs. 47,848 losses) and *occurrence* (17,319 vs. 13,280) tree reconstructions (Figure 1.4B). The gain-to-loss ratios were 1.69 and 1.30, respectively, indicating an increase of 69% and 30% in gains relative to losses. Relative differences in the numbers of gains (red) versus losses (blue) suggested that gains increased with the progression of evolutionary time in both reconstructions (read below).

We note that different evolutionary processes may be responsible for shaping the proteomes in individual superkingdoms. For example, the origin of Archaea has been linked to genome reduction events [20,86], while HGT is believed to have played an important role in the evolution of bacterial species [25]. In contrast, eukaryal proteomes harbor an increased number of novel domain architectures that are a result of gene duplication and rearrangement events [6,43]. Therefore, to eliminate any biases resulting from the effects of superkingdoms in the global analysis (Figure 1.4B), we recalculated the history of character changes on the pruned superkingdom tress recovered earlier (Figure 1.4C).

For *abundance* reconstructions, the exercise supported earlier results where the number of gains was significantly higher than the corresponding number of losses for Archaea (4,616 vs. 2,009), Bacteria (36,606 vs. 20,196), and Eukarya (40,515 vs. 25,036) (Figure 1.4C: *abundance*). The overall gain to loss ratios decreased from 2.30 in Archaea to 1.81 in Bacteria and 1.62 in Eukarya (Figure 1.4C: *abundance*). The increased gain-to-loss ratios in akaryotic microbial

species are remarkable; it implies that the rate of gene discovery in akaryotic microbes (by *de novo* creation, gene duplication, acquisition by HGT and/or recruitment) is higher than the rate in eukaryotes. This tendency in microbial species could be a novel 'collective' persistence strategy to compensate for their economical proteomes. For histograms representing *occurrence* models, global gain-to-loss ratios decreased in the order, Archaea > Bacteria > Eukarya (Figure 1.4C: *occurrence*). Remarkably, the ratio in Eukarya dropped below 1 indicating prevalence of domain loss events relative to gains. This result supports recent studies that have proposed the evolution of newly emerging eukaryal phyla via genome reduction [87].

Accumulation of gains and losses in evolutionary time

When partitioned into the *early*, *intermediate*, and *late* evolutionary epochs, the gain-to-loss ratios exhibited an approximately linear trend towards increasing gains (Figure 1.5). For *abundance*, the ratios increased from 1.32 in the *early* epoch to 1.45 in the *intermediate* epoch and 1.96 in the *late* evolutionary epoch. Similar trends were also observed for *occurrence*, with calculated ratios of 0.61, 0.97, and 1.68, respectively (Figure 1.5A). In fact, both gains and losses increased linearly with evolutionary time in all reconstructions. However, accumulation of gains overshadowed the number of losses (Figure 1.5). Remarkably, the *occurrence* model suggested predominant losses in the first two phases of evolution (0.61 and 0.97) that were compensated by significantly higher amounts of gains (1.68) in the *late* epoch. In contrast, *abundance* failed to illustrate this effect and indicated overwhelming gains in all evolutionary epochs.

When looking at the individual epochs for pruned trees (Figure 1.5B), we noticed that the rate of domain gain increased with time (as before) (Figure 1.5A). However, the ratios in the initial two evolutionary epochs were considerably higher in Archaea for both the *abundance* and *occurrence* models. For example, Archaea exhibited gain-to-loss ratios of 2.06 and 2.14, in comparison to 1.26 and 1.39 in Bacteria, and 1.55 and 1.67 in Eukarya for *early* and *intermediate* evolutionary epochs (Figure 1.5B:*abundance*). In contrast, Bacteria exhibited an overwhelming gain-to-loss ratio of 2.88 in comparison to 2.67 in Archaea and 1.61 in Eukarya, in the *late* evolutionary epoch. Overall, the gain-to-loss ratios increased with evolutionary time in all superkingdoms with the sole exception of Eukarya that had a lower ratio in the *late* (1.61) compared to the *intermediate* (1.67) epoch (Figure 1.5B:*abundance*).

Results based on *occurrence* indicated similar trends but with (relatively) more balanced gain-to-loss ratios and still highlighted the abundance of domain gains in evolution. The individual ratios were 1.42, 1.66, and 2.44 in Archaea, 0.60, 0.91, and 2.61 in Bacteria, and 0.51, 0.95, and 0.95 in Eukarya (Figure 1.5B:*occurrence*). Both Bacteria and Eukarya showed increased levels of ancient domain loss. However, Bacteria offset this decrease by engaging in massive gain events during the *late* evolutionary epoch (ratio of 2.61). In contrast, Eukarya exhibited an even exchange between FF gain and loss events (ratio = 0.95) in both the *intermediate* and *late* epochs. *Occurrence* results also supported the evolution of Eukarya by gene loss, which is in line with recently published analyses [23]. *Abundance* also indicated this drop in gene discovery rate for recent domains in Eukarya. However, the drop appears to be compensated by increased duplications of other domains that lead to an increase in the overall number of domains that are gained (Figure 1.5B:*abundance*). This apparent discrepancy can be explained by the power of both models in depicting true evolutionary relationships between organisms. *Abundance* accounts for a number of evolutionary processes such as HGT, gene duplication, and gene rearrangements while *occurrence* merely describes presence and absence of FFs and because of its more 'global' nature fails to illustrate a complete evolutionary picture (see read below).

Effect of unequal sampling of proteomes

To test whether unequal sampling of proteomes per superkingdom was contributing any bias to the calculations of domain gains and losses, we extracted 100 random samples of 34 proteomes each from the three superkingdoms and generated 100 random trees. From each of the random trees, we recalculated the gain-to-loss ratios using both *abundance* and *occurrence* models (Figure 1.6). Random and equal sampling supported the overall conclusion that gains were overwhelming during the evolution of domain repertoires (Figure 1.6). The median ratios for random trees were 2.47 in Archaea, 2.35 in Eukarya, and 2.34 in Bacteria for *abundance* reconstructions (Figure 1.6A). In comparison, the ratios decreased from 2.11 in Archaea to 1.93 in Bacteria and 1.11 in Eukarya for *occurrence* reconstructions (Figure 1.6B). Based on the results of random and equal sampling, we safely conclude that the gain of domains in proteomes is a universal process that occurs in all three superkingdoms of life. Moreover, the gain-to-loss ratios increase with time (Figure 1.5) and their effects are directly responsible for evolutionary adaptations in superkingdoms. We also propose that using *abundance* increases the reliability of

the phylogenomic model and accounts for many important evolutionary events, a feat that is not possible when studying *occurrence*.

GO Enrichment analysis

We identified FFs that were gained (i.e. net sum of gains and losses was above 0) and lost (net sum below 0) directly from the pruned superkingdom trees. To eliminate any redundancy, we only kept FFs that were gained (or lost) in both *abundance* and *occurrence* reconstructions and excluded those where both methods disagreed. Using this stringent criterion, we classified a total of 368 archaeal FFs as gained and 40 as lost. In comparison, Bacteria and Eukarya gained 892 and 633 FFs, respectively, while they lost only 148 and 164 FFs. Both gained and lost FFs for each superkingdom were provided as input to the online dcGO resource [56,57] to retrieve the highly specific and significantly enriched biological process GO terms (see Methods).

For FFs that were gained, a total of six GO terms were significantly enriched in archaeal proteomes representing biological processes involved in the biosynthesis of nucleotides and metabolism, such as 'tricarboxylic acid cycle [GO:0006099]', 'pyruvate metabolic process [GO:0006090]', 'acyl-CoA metabolic process [GO:0006637]', 'thioester biosynthetic process [GO:0035384]', 'purine nucleobase metabolic process [GO:0006144]', and 'pyrimidine nucleoside metabolic process [GO:0006213]' (Table 1.4). In comparison, only one biological process in Bacteria ('polysaccharide catabolic process [GO:0000272]') and thirty-seven in Eukarya were significantly enriched (Table 1.4). While, the bacterial GO term corresponded to metabolic roles (similar to Archaea), eukaryal functions encompassed a diverse range of processes including 'sex determination [GO:0007530]', regulatory [GO:0044089] and immunological roles [GO:0046634], functions related to the development of mammary glands [GO:0061180], and others (Table 1.4). Finally, none of the archaeal or eukaryal lost FFs was significantly associated with any of the highly specific biological process GO terms, indicating that loss of FFs in these two superkingdoms occurred without any functional constraint. In contrast, two biological processes were predicted to be lost from Bacteria including, 'cellular modified amino acid biosynthetic process [GO:0042398]', and 'pyrimidine-containing compound biosynthetic process [GO:0072528]' (Table 1.5).

Discussion

Evolutionary patterns

We report the evolutionary dynamics of gain and loss events of protein domain FFs in hundreds of free-living organisms belonging to the three cellular superkingdoms. Structural phylogenomic methods were used to reconstruct ToLs from genomic *abundance* and *occurrence* of FF domains in proteomes. Standard character reconstruction techniques were then used to trace domain gain and loss events along the branches of the universal trees. Finally, molecular functions and biological processes of FFs were studied using traditional resources. The exercise revealed remarkable patterns:

(1) Domain gains outnumbered losses throughout evolution. The tracing of character state changes along the branches of ToLs revealed that both domain gain and loss were frequent outcomes in proteome evolution. However, a global trend of gains was pervasive along the entire evolutionary timeline and in all superkingdoms (Figures 1.4-1.6). Remarkably, the gain-to-loss ratios increased with the progression of evolutionary time. However, the rates of domain discovery varied considerably among superkingdoms. To our knowledge, this is the first exercise that has studied gain-and-loss dynamics on a global scale by subjecting all organismal lineages in a ToL to character state reconstruction analysis. Domain gain can lead to interesting evolutionary outcomes. First, it increases the domain repertoire of cells and enhances the persistence strategies of living organisms. Second, the process allows acquisition of novel functions and ensures the availability of more domains for use in the combinatorial interplay that is responsible for the generation of novel domain architectures. In contrast, domain loss is important for changes from free-living to parasitic or symbiotic lifestyles [22] that lead to highly reduced genomes [21].

(2) Secondary evolutionary adaptations are ongoing in superkingdoms. Modeling of FF gain and loss events in proteomes revealed that microbial superkingdoms, especially Archaea, had the highest rates of domain gains (Figures 1.4-1.6). This finding and the fact that the majority of the informational FFs unique to the AE taxonomic group (Table 1.2) were late additions ($nd \geq 0.55$) to the FF repertoires point to another interesting evolutionary adaptation of Archaea: the late discovery and sharing of FFs with other superkingdoms (especially Eukarya) to compensate for the initial evolutionary reductive trend. This secondary archaeal adaptation to offset ancient genome reduction events and the proteomic trends towards economy may also be

occurring in Bacteria (albeit at lower degree), which also exhibited higher levels of gene discovery. In contrast, eukaryal species favored the reuse of already existing domains rather than engaging exclusively in novel domain discovery. Thus, akaryotic microbes persist by fostering trends towards economy while eukaryotic species favor patterns of more flexibility and robustness. However, the low robustness of archaeal species is intriguing and demands an explanation. Archaea are characterized by their preferences for extreme environmental niches (e.g. thermophilic and halophilic environments), a factor intuitively responsible for increased robustness in cells. However, robustness is associated with an organism's ability to respond to changing environmental conditions [64]. Both Bacteria and Eukarya are more diverse in this regard and interact with a diverse range of temperatures, moistures, and climates. In comparison, Archaea are more restricted in terms of their environmental niches and do not generally face varied climatic conditions. In light of these observations, our finding that robustness in cells increased in the order, Archaea, Bacteria, and Eukarya is intuitively well supported.

(3) Functional annotations of timelines revealed differential enrichment of molecular functions in superkingdoms. Annotations of the molecular functions of FFs highlighted the abundance of metabolic and informational domains in proteomes (Figure 1.2A), supporting previous studies [21]. Informational FFs were significantly over-represented in the AE taxonomic group and appeared during the *late* evolutionary epoch. This suggested that both Archaea and Eukarya work with a very similar apparatus for decoding their genetic information, which is different from Bacteria. However, as we explained above, all these innovations occurred in the *late* epoch ($nd > 0.55$), highlighting ongoing secondary adaptations in the superkingdoms. In comparison, the BE taxonomic group was enriched in metabolic FFs (Figure 1.2A). This toolkit was probably acquired via HGT during endosymbiosis of primordial microbes rich in diverse metabolic functions (read below).

The enrichment of biological processes in superkingdoms revealed that akaryotes gained and lost metabolic capabilities during the course of evolution (Tables 1.4 and 1.5), while eukaryotes gained a significant number of functionalities involved in the diversification of eukaryal lineages such as the development of mammary glands, compound eye development, enhanced regulatory roles, and sex determination (Table 1.4). All these processes reflect relatively recent evolutionary innovations in the eukaryal superkingdom suggesting that while the overall rate of innovation was lowest in Eukarya; it was directed towards discovering

important functions responsible for the diversification of eukaryal phyla and kingdoms (e.g. appearance of mammals) from the last common eukaryotic ancestor. However, we caution that the significantly enriched GO terms (Tables 1.4 and 1.5) only represent a subset of FFs (i.e. those corresponding to gains and losses) from the entire FF repertoires in superkingdoms. Thus they do not reflect the entire toolkit of biological processes that are expected to occur in the living organisms and should be interpreted with limited scope.

(4) Early origin of Archaea by genomic streamlining. ToLs generated from genomic *abundance* and *occurrence* counts were rooted paraphyletically in Archaea, a result that disagrees with the canonical rooting of Bacteria recovered from 16S rRNA and ancient paralogous gene sequence trees [74,81]. The archaeal rooting of the universal tree is supported by a number of previous studies involving more conserved phylogenetic characters describing the structure and function of both proteins and RNA molecules [38,84,85,88,89]. We have previously argued that trees built from protein domain structure (i.e. FSFs and FFs) are robust against a number of problems that complicate phylogenetic analysis of gene sequences [37]. First, gene sequences are prone to high mutation rates [90] and are far less conserved than protein domain structures [20]. Second, computation of a reliable sequence alignment is a painstaking process and involves manual editing [91]. Third, alignment forces unnecessary assumptions about inapplicable characters such as insertion/deletions [92,93]. Fourth, sequence sites in genes interact with each other to form secondary structures and domain regions and consequently do not change independently from each other [94-96]. Thus each nucleotide cannot be considered an independent character in phylogenetic analyses [37]. These and other shortcomings (see [37]) limit and reduce the reliability of sequence-based methods and cast doubt on statements of deep phylogeny such as the canonical rooting of the ToL.

Moreover, the 16S rRNA gene that is considered the gold standard for phylogenetic analysis only represents one component of the ribosome, a central macromolecular complex that holds at least two other rRNA components and many structural proteins with varying evolutionary histories [97]. Thus, trees built from rRNA genes can only provide a glimpse of the evolutionary history of the ribosome and not the entire organismal systems that are made up of many biological parts. Our approach is advantageous in this regard as it studies the evolution of systems (organisms) using their component parts (entire domain repertoire) and provides a global perspective. Finally, our approach does not require computation of any alignment and does not

violate the assumption of character independence, as each SCOP FF is an independent evolutionary unit [37].

The distribution of FF domains in superkingdoms also showed that both the numbers of unique and shared FFs were lowest in Archaea. For example, the number of FFs shared between Bacteria and Eukarya was considerably higher than those shared with Archaea (BE = 412 vs. AB = 90 and AE = 40) (Figure 1.1A). Without any formal phylogenetic analysis, it is evident from the patterns of use and sharing of domain structures in Venn diagrams (Figure 1.1A) and the 3D-plots describing persistence strategies (Figure 1.1D), that Archaea represents the simplest form of cellular life. The smaller FF domain repertoires in archaeal species could be an outcome of one of two possible events: (i) Archaea evolved by gradual loss of ancestral genes (via genome reduction) when nascent lineages delimited the emergence of the first superkingdom of life, or (ii) Both Bacteria and Eukarya gained a significant number of FFs later in evolution (after diverging from Archaea), while the archaeal superkingdom persisted in its path of economy. While both of these scenarios point to an early origin of the archaeal superkingdom, our data and previous results [27] are more compatible with the former event.

We have previously argued that the complete absence of an 'ancient' fold (FF or FSF) in one superkingdom more likely represents a loss event in that superkingdom rather than simultaneous gains of the same fold in other superkingdoms [20]. In other words, the probability of one group loosing a structure is higher than two groups acquiring the same structure at the same time. Under this probabilistic model, the appearance of the BE taxonomic group at *nd* = 0.15 represents a fundamental evolutionary event of complete loss of ancient FFs in the archaeal superkingdom (Figure 1.1B). Our data confirm that the first FF to be lost from Archaea was the 'Heat-shock protein, HSP90, N-terminal domain', which is highly conserved in bacterial and eukaryotic species but completely absent in Archaea [63]. Lack of HSP90 chaperones in Archaea is intriguing and merits future exploration of how protein-folding mechanisms work in extremophiles. A recent analysis of FSF domains [14] also confirmed that Archaea evolved by genome reduction and that this process started very early in evolution. In that study, the distribution (*f-value*) of 1,739 FSFs in 70 archaeal proteomes revealed that many of the ancient folds were completely absent in archaeal species. This hypothesis is strengthened by our data of minimal sharing of FFs in archaeal taxonomic groups (Figure 1.1A) and the appearance of taxonomic groups (Figure 1.1B), suggesting an early evolutionary split of Archaea (Figure 1.3).

In light of these observations, our finding that the origin of diversified cellular life lies in thermophilic archaeal species (Figure 1.3) is a significant outcome that is supported by sound methodological and evolutionary considerations.

(5) A canonical pattern of superkingdom diversification embeds the likely endosymbiotic origin of eukaryotes: FF distributions in the evolutionary timeline of domain appearance revealed that Archaea was the first superkingdom to materialize by selective loss of domain structures at the end of the *early* epoch of evolution (Figure 1.1B and 1.2). Remarkably, however, the appearance of superkingdom-specific domains followed an order that matches the canonical pattern of early rise of Bacteria during the *intermediate* epoch and joint rise of diversified Archaea and Eukarya at the start of the *late* epoch. Thus, the primordial stem line, which was already structurally and functionally quite complex, generated organismal biodiversity first by streamlining the structural make up in Archaea (at $nd = 0.15$), then by generating novelty in Bacteria ($nd = 0.26$), and finally by generating novelty and co-opting bacterial lineages as organelles in Eukarya ($nd < 0.55$). The eukaryotic group was able to deploy massive structural and functional innovation (despite concomitant streamlining), which we show spread through eukaryotic lineages at high frequency (Figure 1.2C). Tendencies of flexibility and robustness of this kind were neither deployed by the akaryotic superkingdoms that preceded Eukarya nor by superkingdom-specific diversification of the archaeal domain repertoires that coincided with its rise.

Our data is thus incompatible with fusion scenarios between akaryotic cells that are used to explain the origin of eukaryotes [66,67,69,98], which have been criticized previously [70,76-79] and are not supported by comparative proteomics analysis [78]. They also fail to explain the presence of bacterial-like lipids in eukaryotes, especially if the partner cells were archaeons and bacteria (e.g. [67]). Moreover, no known mechanism of akaryotic engulfment exists, no extant bacterium is known to enter or survive inside archaeal organisms, and cellular fusion is incompatible with archaeal cell biology. In contrast, there is considerable evidence supporting the endosymbiotic origins of eukaryotic organelles. It is highly likely that mitochondria developed from the SAR11 clade of marine bacteria, a sister group to the *Ricketsialles* [99]. There is also considerable evidence in support of eukaryotic mechanisms of phagocytosis that would enable microbial engulfment of organelle ancestors [100]. The question however relates to the defining event of eukaryal diversification. Our timelines indicate the presence of an ancestral

proto-eukaryotic stem lineage that was structurally and metabolically quite advanced. This lineage already produced superkingdoms Archaea and Bacteria by genomic streamlining, which was likely triggered by a host of selective pressures, including the escape from viruses and phagotrophs, the need to adapt to extreme environments (Archaea), and exploring the benefits of rapid growth (Bacteria) [70]. The early rise of diversified Bacteria supports the existence of alpha-proteobacterial ancestors of mitochondria before the appearance of diversified eukaryotes 1.6 Gy ago (*nd* = 0.55), as indicated by microfossil evidence and the molecular clock [51,52]. The fact that the first mitochondrial-specific FFs appeared at that time (Table 1.3) boosts the idea of the joint rise of Eukarya and eukaryotic organelles. It is therefore highly likely that the proto-eukaryotic stem line acquired phagotrophic abilities and engulfed an alpha-proteobacterium and other microbes (including archaeons) to trigger the diversification of eukaryotes soon after. This scenario [78] seems most compatible with our timelines and explains the enrichment of metabolic BE domains. A formal test of the phagotrophic proto-eukaryotic ancestor is warranted.

Reliability of our study

How reliable is our study? Both *abundance* and *occurrence* were congruent with respect to the overall tree topologies and general conclusions drawn from the analysis. Both supported the existence of overwhelming gains in evolution. However, discrepancies also existed especially in the numerical differences for the gain-to-loss ratios among superkingdoms. In general, *abundance* (apparently) overestimated gains while *occurrence* underestimated losses. The higher number of gain-to-loss ratios in *abundance* models is an expected outcome as we are accounting for evolutionary processes such as gene duplications, gene rearrangements, and HGT that are known to increase the representation of genes in genomes. Ancient genes have more time to multiply and increase their genomic abundance compared to newly emergent genes. In contrast, *occurrence* merely describes the presence or absence of genes and provides a simplified view of the overall landscape of change. Another explanation is the possible existence of methodological artifacts when dealing with genomic occurrence in parsimony analysis that excludes most of the ancient FFs as non-informative characters, when these are present in all proteomes. Moreover, *occurrence* fails to take into account the weighted contribution of ancient genes to the phylogeny and treats all characters equally. Thus trees built from *abundance* counts are better resolved at their base while trees built from *occurrence* behave poorly in this regard [27]. We emphasize that the focus of this study is to highlight the relative contribution of domain gains and losses in the

evolution of superkingdoms and not to evaluate which methodology is preferable. The finding that domain gains are overwhelming and increase approximately linearly with evolutionary time in both models is remarkable and suggests that the appearance of novel domains is a continuous process (Figures 1.4 and 1.5).

In our phylogenomic model, we rooted ToLs by character absence (i.e. 0) using the Lundberg method. We assumed that proteomes became progressively richer during the course of evolution. However, this implicit assumption did not lead to an increased number of domain gains as character state changes in both forward (e.g. 9 to 22) and reverse (12 to 5) directions were allowed and carried equal weights. Moreover, we evaluated the effects of ToL rooting on the calculations of domain gain and loss statistics by considering outgroup taxa instead of the Lundberg method. Superkingdom trees rooted with outgroup taxa led to similar tree topologies and supported the conclusion of overwhelming gains that we here report (Figure A1). However, we decided to exclude outgroup analysis from this study for two reasons. First, outgroups add an external hypothesis into the model and bias gains and losses by including artificial character changes in the most basal branches leading to outgroup taxa. Second, the selection of the most appropriate outgroups for each superkingdom is a complicated problem and is virtually impossible for the reconstruction of ToLs. However, it would be interesting to study the gain and loss dynamics at different levels of the SCOP hierarchy such as the FSF and F levels of structural abstraction. We expect that patterns reported in this study will remain robust regardless of the SCOP conservation level and will extend the analysis to FSF in a separate publication.

We used MP to search for the best possible tree and described the evolution of 420 free-living proteomes using the entire repertoire of 2,397 FFs as phylogenetic characters. We note that MP is most appropriate (and gives superior performance) for this kind of analysis as it performs better when the characters are evolving under different evolutionary rates [101]. Moreover, rescaling of raw abundance values into 24 possible character states considerably reduces the likelihood of convergent evolution. Reconstructing evolutionary history of species and studying domain emergence and loss patterns is a difficult problem complicated by a number of considerations (e.g. taxa and character sampling, biases introduced by organism lifestyles, ecological niches of organisms, and non-vertical evolutionary processes). We attempted to eliminate these problems by reconstructing whole-genome phylogenies, sampling conserved FF domains as characters, excluding parasitic and facultative parasitic organisms from study, and by

using multistate phylogenetic characters. However, we realize that no method is free from technical and logical artifacts. Our analysis largely depends upon the accuracy of phylogenetic reconstruction methods, current SCOP domain definitions, reliability of function annotation schemes, and literature for organism lifestyle. However, we expect that recovered results will remain robust both with data growth and improvement in available methods and that drastic revisions to existing databases would be unlikely. For that reason we caution the reader to focus on the general trends and main conclusions of the paper (i.e. overwhelming gains and its consequences) rather than the actual numbers and discrepancies between the phylogenomic methods. Quantifying gain and loss events on a global scale is a difficult problem and our work lays foundations for more and improved studies in the future.

Conclusions

We propose that grouping of protein domains into FFs provides a reliable character for a global evolutionary analysis that involves large number of proteomes. Domain FFs are both sufficiently conserved and informative to explore the many branches on the ToLs. The age and distribution of FFs in organismal groups is biased and carries the power to unfold superkingdom history and explain important structural and functional differences among superkingdoms. Based on our data, we propose the primacy of domain gains over losses over the entire evolutionary period, ongoing evolutionary adaptations in akaryotic microbes, evolution of emerging eukaryotic species by domain loss, an early origin for Archaea, and endosymbiosis leading to mitochondria as a crucial event in eukaryote diversification. Each of these conclusions is important for reconstructing the evolutionary past and predicting evolutionary events in the future.

Figures

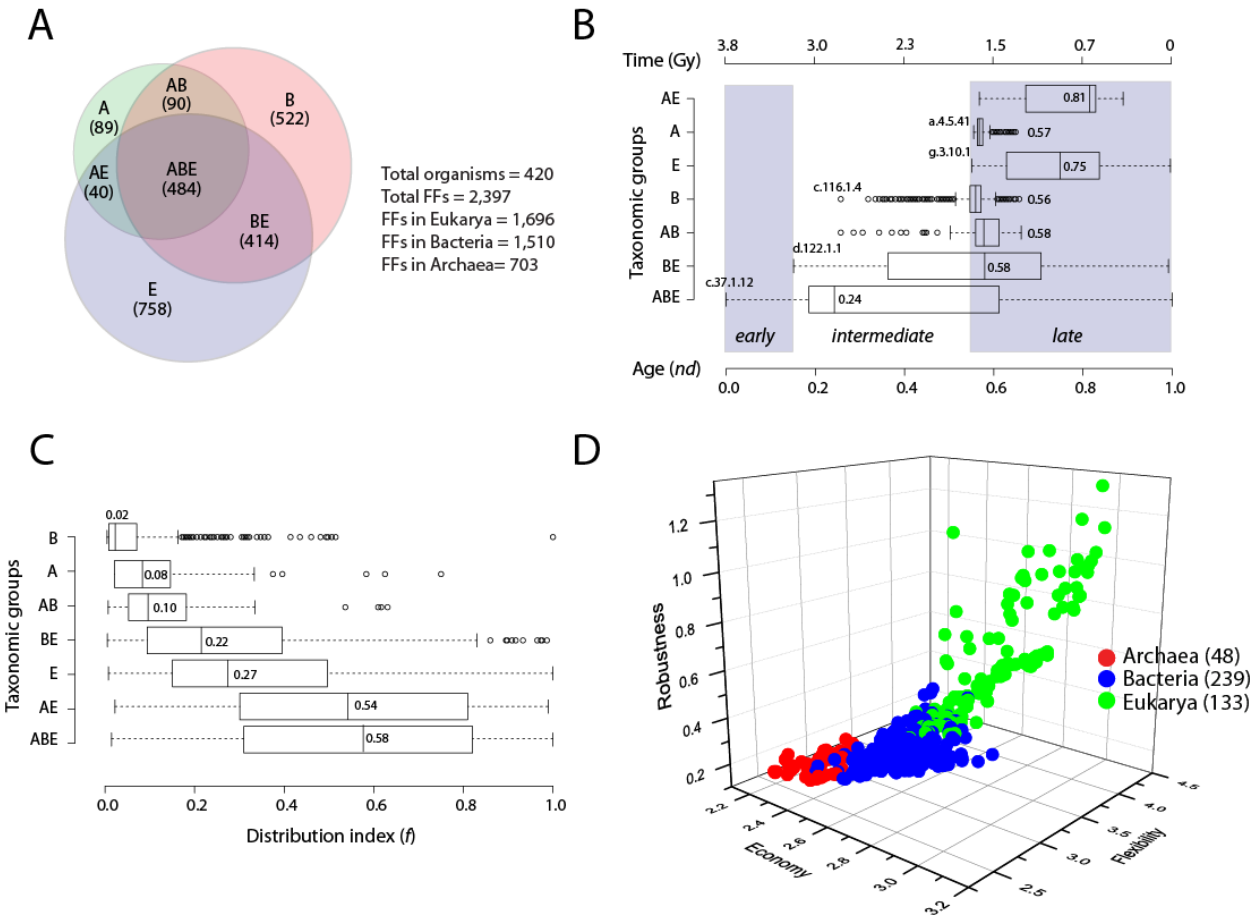

Figure 1.1 Evolutionary dynamics of FFs and organismal persistence strategies. A) A Venn diagram describes the distribution of FFs in the seven taxonomic groups (reproduced from [27]). **B)** Boxplots represent the distributions of domain ages (*nd*) for each taxonomic group. Numbers within each distribution indicate group medians, hollow circles the outliers, while the shaded regions identify important evolutionary epochs. Geological time (Gy) was inferred from a molecular clock of protein folds [51,52]. FFs were identified by SCOP *css:* c.37.1.12, ABC transporter ATPase domain-like; d.122.1.1, Heat shock protein 90, HSP 90, N-terminal domain; c.116.1.4, tRNA(m1G37)-methyltransferase TrmD; g.3.10.1, Colipase-like; a.4.5.41, Transcription factor E/IIe-alpha, N-terminal domain. **C)** Boxplots highlight the distribution (*f*-value) of FF domains in the proteomes of each taxonomic group. Numbers within each distribution indicate group medians. Hollow circles represent outliers. **D)** A 3D scatter plot describes the persistence strategies of Archaea (red), Bacteria (blue), and Eukarya (green). All axes are in logarithmic scale. Numbers in parenthesis indicate total number of proteomes available for study in each superkingdom.

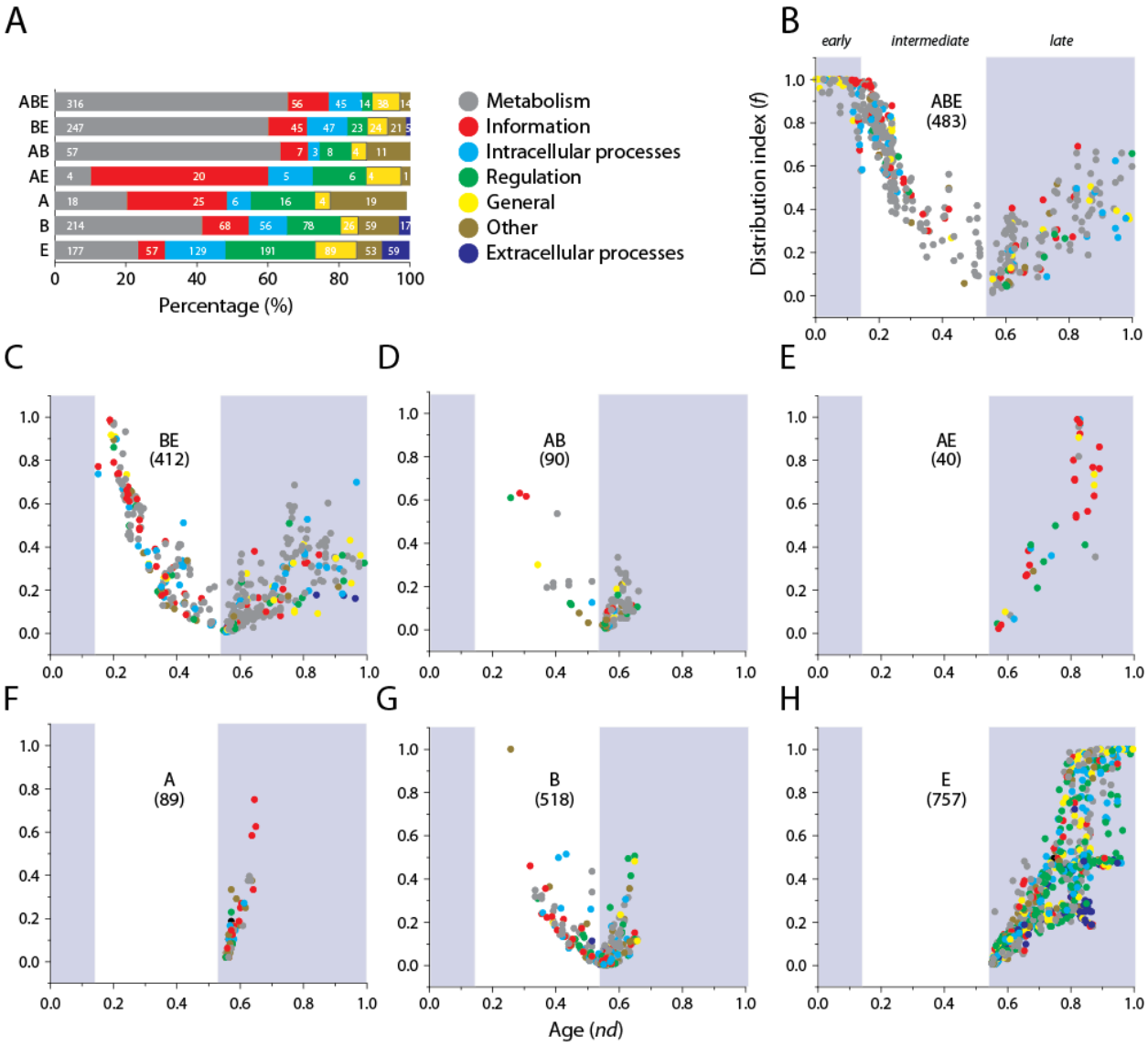

Figure 1.2 Functional annotation of FF domains. A) Stacked bar plots describe the distribution of molecular functions in each of the seven taxonomic groups. The size of each bar is proportional to the percentage of FF domains in each functional category, while the numbers indicate total counts of FFs annotated in that category. **B-H)** Scatter plots illustrate the emergence of molecular functions in taxonomic groups. The *x*-axes represent evolutionary time (*nd*) while the *y*-axes indicates the distribution index (*f*-value) of FFs. Evolutionary epochs identified as previously. Numbers in parenthesis indicate total number of FF domains in each taxonomic group for which SUPERFAMILY functional annotations (based on SCOP 1.73) were available.

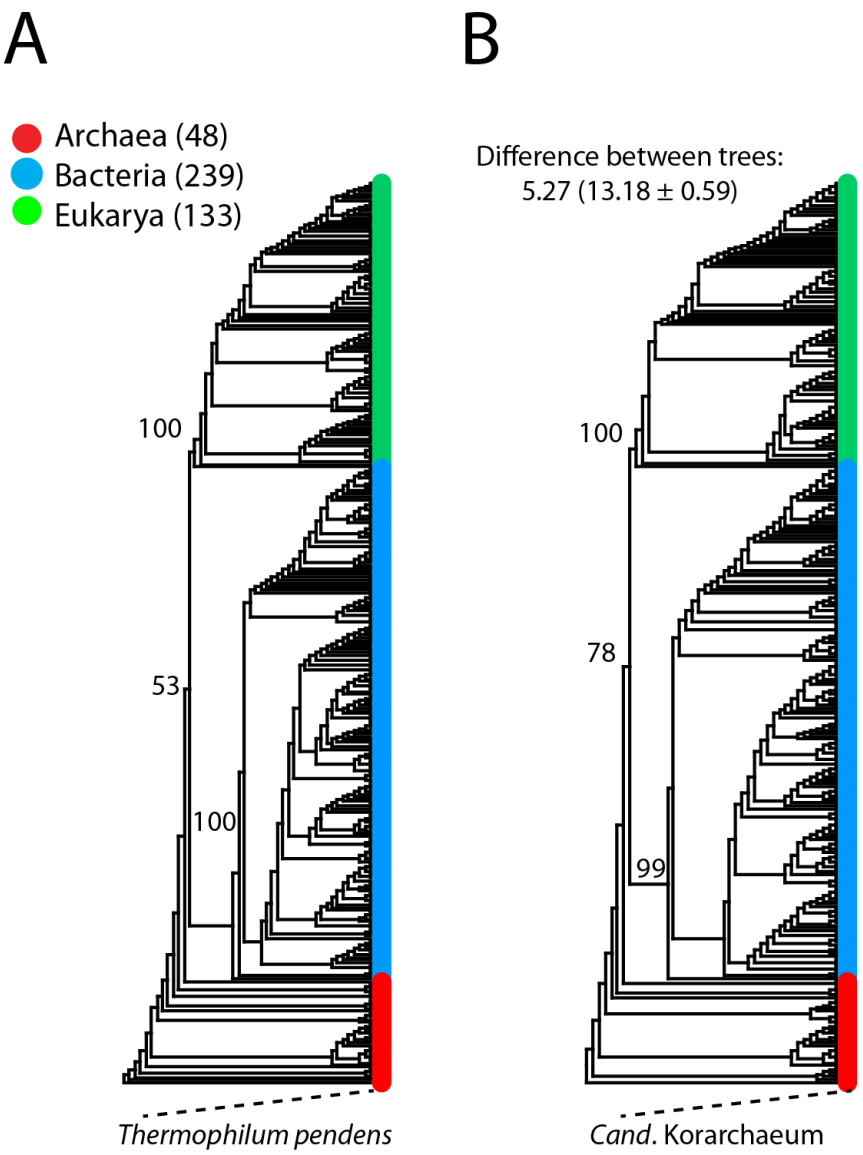

Figure 1.3 Phylogenomic patterns in the three superkingdoms. A) A ToL reconstructed from the genomic abundance counts of 2,397 FF domains (2,262 parsimony informative, tree length = 128,752, RI = 0.76, g_1 = -0.33) describing the evolution of 420 free-living organisms. Values on branches indicate bootstrap support values. Taxa were colored red for Archaea, blue for Bacteria and green for Eukarya. **B)** A ToL reconstructed from the presence/absence of 2,397 FF domains (2,249 parsimony informative, tree length = 30,599, RI = 0.79, g_1 = -0.28) describing the evolution of 420 free-living organisms. Values on branches indicate bootstrap support values. Taxa colored as in **(A)**. Difference between trees was calculated using the nodal module of TOPD/FMTS package [50].

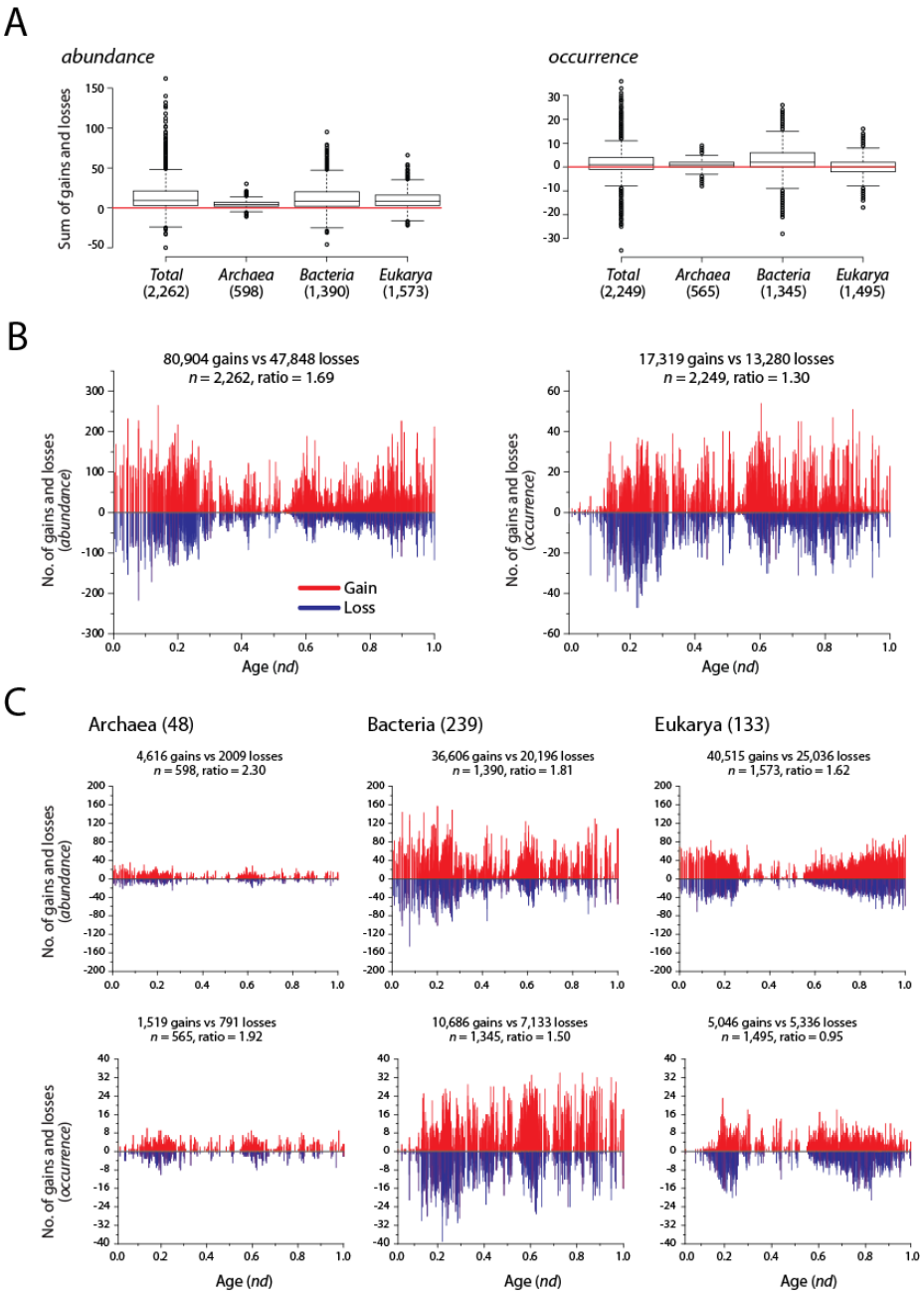

Figure 1.4 Global patterns of FF gain and loss in superkingdoms. A) Sum of gains and losses for each FF domain is represented in boxplots for *Total*, *Archaea*, *Bacteria*, and *Eukarya* reconstructions using *abundance* and *occurrence* models. Numbers in parentheses indicate total number of parsimony informative characters in each analysis. A horizontal red line passes through zero on the *x*-axis. **B)** Histograms comparing the relative counts of gains and losses for each FF domain character, plotted on the *nd* scale. Bars in red and blue indicate gains and losses, respectively. The global gain-to-loss ratios are listed along with the total number of gain and loss events and gain-to-loss ratios. *n* is the number of parsimony informative characters in each analysis. **C)** Histograms comparing the distribution of FF gain and losse in *Archaea*, *Bacteria* and *Eukarya*. Bars in red and blue indicate gains and losses, respectively. The *x*-axis indicates evolutionary time. Numbers in parenthesis indicate total number of proteomes in each dataset.

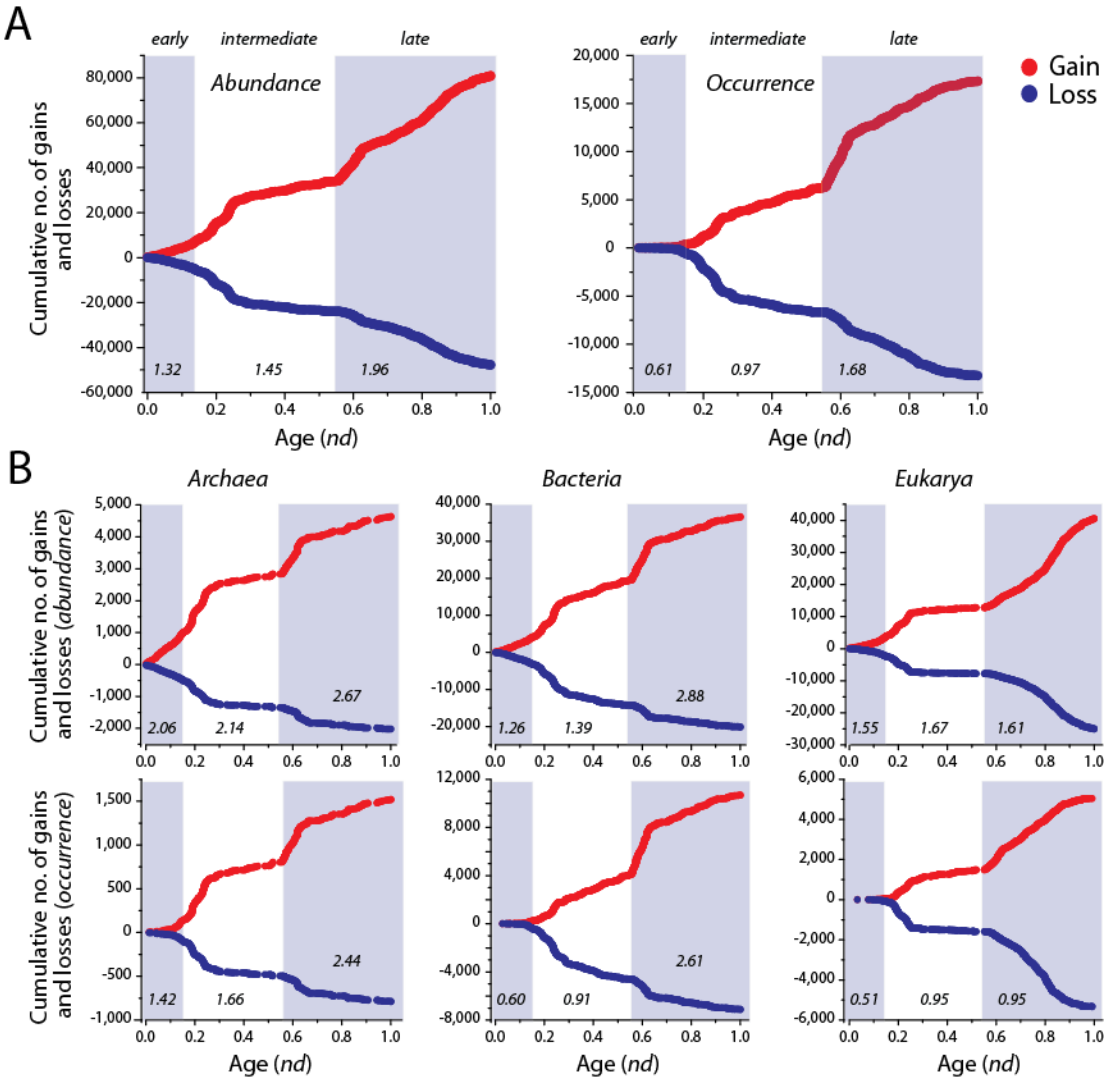

Figure 1.5 Cumulative numbers of gains and losses. Scatter plots reveal an approximately linear trend in the accumulation of FF gains and losses in both the global analysis **(A)** and in individual superkingdoms **(B)**. Gains are identified in red while losses in blue. The three evolutionary epochs are marked with corresponding gain-to-loss ratios in italics.

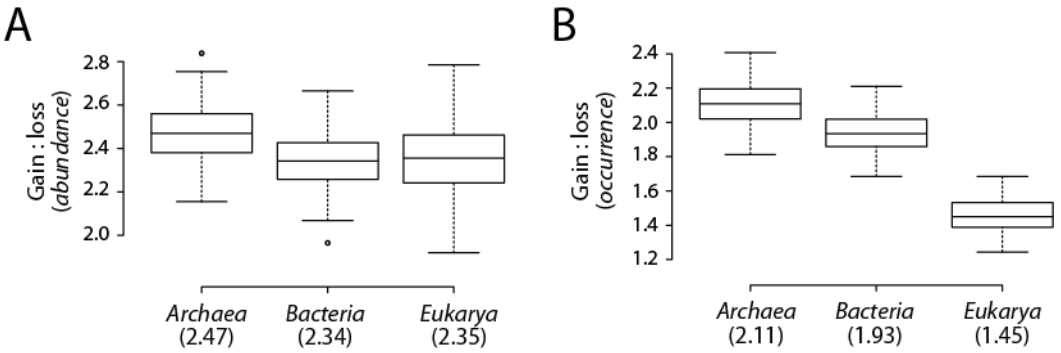

Figure 1.6 Equal sampling of proteomes. Boxplots comparing the distribution of net gains and losses in 100 random phylogenetic trees for *abundance* **(A)** and *occurrence* **(B)**. Numbers in parentheses indicate group median values.

Tables

Table 1.1 Descriptive statistics on the total number of proteomes (*N*), minimum (*min*), maximum (*max*) and median values for raw counts of occurrence, abundance and ratio of FFs in each superkingdom. The superscripts identify individual species.

		Occurrence			**Abundance**			**Ratio**		
Superkingdom	*N*	*Min*	*max*	*median*	*Min*	*Max*	*median*	*min*	*max*	*median*
Archaea	48	174[1]	293[2]	236	264[1]	598[2]	377.50	1.46[3]	2.10[4]	1.64
Bacteria	239	239[5]	824[6]	426	376[5]	1958[7]	883.00	1.52[8]	3.40[9]	1.98
Eukarya	133	364[10]	1089[11]	674	982[12]	19917[13]	2875.00	2.24[12]	20.41[13]	4.04

[1]*Staphylothermus marinus,* [2]*Methanosarcina acetovirans,* [3]*Thermoplasma volcanium,* [4]*Haloarcula marismortui,* [5]*Dehalococcoides* sp.*,* [6]*Citrobacter koseri,* [7]*Burkholderia xenovorans,* [8]*Nitratiruptor* sp.*,* [9]*Rhodococcus* sp.*,* [10]*Paramecium tetraurelia,* [11]*Homo sapiens,* [12]*Malassezia globosa,* [13]*Takifugu rubripes*.

Table 1.2 Names, SCOP *css,* and *f*-value of informational FF domains present in the AE taxonomic group. FFs are sorted by *f*-value in a descending manner.

No.	Name	SCOP *css*	*f*-value
1	L30e/L7ae ribosomal proteins	d.79.3.1	0.99
2	Ribosomal protein L3	b.43.3.2	0.99
3	L15e family	d.12.1.2	0.97
4	Ribosomal protein L10e family	d.41.4.1	0.92
5	TATA-box binding protein (TBP), C-terminal domain family	d.129.1.1	0.86
6	N-terminal domain of eukaryotic peptide chain release factor subunit 1, ERF1 family	d.91.1.1	0.80
7	DNA polymerase processivity factor	d.131.1.2	0.77
8	Sm motif of small nuclear ribonucleoproteins, SNRNP family	b.38.1.1	0.76
9	Eukaryotic DNA topoisomerase I, N-terminal DNA-binding fragment family	e.15.1.1	0.71
10	Eukaryotic DNA topoisomerase I, catalytic core family	d.163.1.2	0.71
11	eEF-1beta-like family	d.58.12.1	0.64
12	Eukaryotic type KH-domain (KH-domain type I) family	d.51.1.1	0.56
13	RNA polymerase subunit RPB10 family	a.4.11.1	0.55
14	RPB5 family	d.78.1.1	0.55
15	Ribosomal protein L19 (L19e) family	a.94.1.1	0.38
16	Ribosomal protein L13 family	c.21.1.1	0.31
17	DNA replication initiator (cdc21/cdc54) N-terminal domain family	b.40.4.11	0.27
18	Initiation factor IF2/eIF5B, domain 3 family	c.20.1.1	0.27
19	AlaX-like family	d.67.1.2	0.04
20	VMA1-derived endonuclease (VDE) PI-SceI protein	d.95.2.2	0.02

Table 1.3 Names, SCOP Id and *css*, and evolutionary age (*nd*) of FFs that were identified by keyword search 'Mitochondria' on the dataset of 2,397 FF domains. FFs are sorted by *nd* value in an ascending manner.

SCOP Id	SCOP *css*	Description	Age (*nd*)
69533	c.55.3.7	Mitochondrial resolvase ydc2 catalytic domain	0.55
81422	f.23.5.1	Mitochondrial cytochrome c oxidase subunit VIIb	0.59
81426	f.23.6.1	Mitochondrial cytochrome c oxidase subunit VIIc (aka VIIIa)	0.59
81418	f.23.4.1	Mitochondrial cytochrome c oxidase subunit VIIa	0.63
111358	f.45.1.1	Mitochondrial ATP synthase coupling factor 6	0.64
81414	f.23.3.1	Mitochondrial cytochrome c oxidase subunit VIc	0.65
54530	d.25.1.1	Mitochondrial glycoprotein MAM33-like	0.71
81410	f.23.2.1	Mitochondrial cytochrome c oxidase subunit VIa	0.71
81405	f.23.1.1	Mitochondrial cytochrome c oxidase subunit IV	0.73
47158	a.23.4.1	Mitochondrial import receptor subunit Tom20	0.74
103507	f.42.1.1	Mitochondrial carrier	0.96

Table 1.4 GO Ids, names and *P*-values for highly specific biological processes that were significantly associated ($FDR < 10^{-2}$) with FF gains in Archaea, Bacteria, and Eukarya.

Superkingdom	No.	GO Id	Biological process	*P*-value
Archaea	1	GO:0006099	tricarboxylic acid cycle	5.38E-06
	2	GO:0006090	pyruvate metabolic process	2.80E-05
	3	GO:0006637	acyl-CoA metabolic process	4.01E-05
	4	GO:0035384	thioester biosynthetic process	3.32E-04
	5	GO:0006144	purine nucleobase metabolic process	5.71E-04
	6	GO:0006213	pyrimidine nucleoside metabolic process	6.38E-04
Bacteria	1	GO:0000272	polysaccharide catabolic process	1.26E-04
Eukarya	1	GO:0045995	regulation of embryonic development	1.44E-06
	2	GO:0051588	regulation of neurotransmitter transport	3.35E-06
	3	GO:0001707	mesoderm formation	7.48E-06
	4	GO:0001649	osteoblast differentiation	1.29E-05
	5	GO:0050870	positive regulation of T cell activation	3.45E-05
	6	GO:0030336	negative regulation of cell migration	8.88E-05
	7	GO:0048017	inositol lipid-mediated signaling	1.05E-04
	8	GO:0000165	MAPK cascade	1.16E-04
	9	GO:0051291	protein heterooligomerization	1.21E-04
	10	GO:0046620	regulation of organ growth	2.43E-04
	11	GO:0051099	positive regulation of binding	3.00E-04
	12	GO:0043627	response to estrogen stimulus	3.00E-04
	13	GO:0051216	cartilage development	2.96E-04
	14	GO:0061180	mammary gland epithelium development	2.96E-04
	15	GO:0030856	regulation of epithelial cell differentiation	3.02E-04
	16	GO:0051703	intraspecies interaction between organisms	4.13E-04
	17	GO:0032496	response to lipopolysaccharide	4.07E-04
	18	GO:0032946	positive regulation of mononuclear cell proliferation	5.10E-04
	19	GO:0032869	cellular response to insulin stimulus	5.10E-04
	20	GO:0045580	regulation of T cell differentiation	6.59E-04
	21	GO:0060191	regulation of lipase activity	6.59E-04
	22	GO:0045834	positive regulation of lipid metabolic process	6.59E-04
	23	GO:0050673	epithelial cell proliferation	6.59E-04
	24	GO:0021761	limbic system development	8.39E-04
	25	GO:0046634	regulation of alpha-beta T cell activation	8.39E-04
	26	GO:0045667	regulation of osteoblast differentiation	8.39E-04
	27	GO:0007492	endoderm development	8.39E-04
	28	GO:0044089	positive regulation of cellular component biogenesis	1.04E-03
	29	GO:0007530	sex determination	1.04E-03
	30	GO:0045598	regulation of fat cell differentiation	1.04E-03
	31	GO:0051057	positive regulation of small GTPase mediated signal transduction	1.25E-03
	32	GO:0048749	compound eye development	1.31E-03
	33	GO:0050773	regulation of dendrite development	1.31E-03
	34	GO:0060443	mammary gland morphogenesis	1.31E-03
	35	GO:2001236	regulation of extrinsic apoptotic signaling pathway	1.31E-03

Table 1.4 (contd.)

Superkingdom	No.	GO Id	Biological process	*P*-value
	36	GO:0016055	Wnt receptor signaling pathway	1.31E-03
	37	GO:0046488	phosphatidylinositol metabolic process	1.31E-03

Table 1.5 GO Ids, names and *P*-values for highly specific biological processes that were significantly associated ($FDR < 10^{-2}$) with FF loss in Bacteria. No significant biological process was lost in either Archaea or Eukarya.

Superkingdom	No.	GO Id	Biological process	*P*-value
Bacteria	1	GO:0042398	cellular modified amino acid biosynthetic process	3.10E-04
	2	GO:0072528	pyrimidine-containing compound biosynthetic process	3.10E-04

CHAPTER 2: A TREE OF CELLULAR LIFE INFERRED FROM A GENOMIC CENSUS OF MOLECULAR FUNCTIONS[2]

Introduction

Evolutionary genomics embraces the study of phylogenomic relationships between organisms at global scale. Phylogenomic trees are non-reticulated network representations of molecular evolution with branches, nodes and leaves (taxa) describing change in features of evolving genomes. Prior to molecular biology, phylogenetics was mostly restricted to the study of morphological, biochemical and physiological data. This data did not allow systematic comparison across lineages and made impossible the elucidation of the deep evolutionary relationships of organisms belonging to the three superkingdoms of life (reviewed in [102]). Advances in molecular biology enabled the use of molecular data for phylogenetic tree reconstruction, including the sequence [103], order [104] and content [105] of genes in genomes, and the atomic structural annotation of gene products [28,89,106]. This led to significant evolutionary discoveries such as recognition of Archaea as the third domain of life [74,107], illustration of reductive trends in the genomes of cellular organisms [20,86], and the genetically simple but functionally complex make up of the last universal common ancestor (LUCA) of life [38].

Reconstructing phylogenetic trees from protein and nucleic acid sequences has become common practice. However, the use of sequence information may not be appropriate for studying deep phylogenetic relationships. In fact, mutation, recombination and gene duplication of molecular sequences occurs at relatively fast pace [108-110]. This dynamics leads to mutational saturation and paralogy, important processes that limit the validity of phylogenetic analysis to low taxonomy-level snapshots of recent evolutionary history. Although a few highly conserved orthologous genes are still available for reconstructing global phylogenies of living organisms, including the tree of life (ToL), their information cannot fully resolve relationships that are deep (e.g. polytomies in rRNA trees; [111]). A few recent studies have reconstructed ToLs using protein domain repertoires [28,112], domain interactomes [43] or metabolic information

[2]This chapter has been published as manuscript in *Journal of Molecular Evolution* (see [301]). The final publication is available at http://link.springer.com/article/10.1007%2Fs00239-014-9637-9. Authors are allowed to self-archive the author accepted version.

[113,114]. These new kinds of data are regarded as controlled molecular vocabularies that cover the continuous spectrum of evolutionary conservation. While the new phylogenies resemble traditional classifications, they yield novel insights into the emergence and evolution of cellular life. Here we expand on the idea of reconstructing ToLs from atypical genomic information by producing rooted phylogenies derived directly from the entire repertoires of molecular functions (functionomes).

The Gene Ontology (GO) database describes the functional annotations and relationships of nearly half a million proteins [58]. This information is presented in three separate tree-like structures, in which three root GO terms, molecular function [MF], biological process [BP], and cellular component [CC], descend toward a bottom (terminal) level into a multi-level hierarchy of ontological terms. Each of these tree-like structures represents an independent directed acyclic graph (DAG) where child GO terms can be associated with multiple parents to account for both differing relationships and associations (Figure 2.1). In the case of DAG_{MF}, GO terms at higher levels represent broader functional categories (e.g. catalytic activity) while those at lower levels indicate more specific functional annotations (e.g. ATPase activity) [58,88]. This hierarchical structure is absent from other existing functional classification schemes such as the Cluster of Orthologous (COG) groups [115] and the functional classification of the SUPERFAMILY database [55]. Although the SEED subsystems provide a hierarchy of multiple functional levels that is similar to the GO, the database specializes in bacterial gene annotation [116]. Consequently, the GO is far more comprehensive than existing databases and has been successfully utilized in the past to describe the evolution of modern molecular functions [88].

We note that the GO hierarchy can be analogous to an evolutionary hierarchy where higher-level GO terms may be more ancient while lower-level terms seem more modern [88]. This notion follows the hypothesis that promiscuous functions can serve as evolutionary starting points [117,118], with proteins of ancient origin being promiscuous and serving multiple functions (comparable to higher-level GO terms) and proteins of recent origin carrying more specified functions (comparable to terminal terms). The existence of this link between GO hierarchy and evolution enables sampling GO terms as phylogenetic characters in hundreds of completely sequenced proteomes (which are considered taxa) and studying the evolution of organisms using a new and more biologically controlled vocabulary. One limitation associated with this approach, however, is the possible effect on phylogenetic reconstruction of non-vertical

evolutionary processes, such as convergent evolution and horizontal gene transfer (HGT). Because GO terms are structured as DAGs, there are *many-to-many* relationships between child and parent terms. This promiscuity can complicate attempts of ToL reconstruction. In addition, genes whose specific functions are not known can be directly assigned to higher-level GO terms without lower-level GO annotations [119]. Consequently, a higher-level GO term is the collection of both evolutionary conserved and functionally unidentified genes.

In this study, we thus restricted the analysis to include only GO terms corresponding to the terminal terms of MF (hereinafter simply referred to as GO_{TMF} terms), which are highly specialized and represent the majority of molecular functions of cells [58]. In contrast, BP represents events that are outcomes of molecular activities (e.g. pyrimidine metabolic process) while CC expresses anatomical structures (e.g. ribosome), both of which carry more integrative views and are not as meaningful for evolutionary studies [88]. Experimentally, we sampled organisms from the three superkingdoms and counted the number of times each GO_{TMF} term was present in their functionomes, and transitively, in their associated proteomes (see Methods). These 'genomic abundance' values serve as phylogenetic character states, characterizing the set of functionomes (taxa) that were sampled (Figure 2.1). The methodology is similar to the abundance-based approach used previously to study the evolution of protein domain structures and RNA molecules [14,27,28,51,97] and is far superior to typical sequence-based approaches that are prone to phylogenetic limitations and artifacts, including problems resulting from sequence alignment (e.g. inapplicable characters and indels that make phylogenetic analysis statistically inconsistent [120], mutational saturation, HGT, and violation of assumptions of character independence [37]). Using this new methodology we show that ToLs reconstructed from the genomic census of GO_{TMF} terms carry considerable predictive power in their ability to explain the origin and evolution of cellular life.

Methods

Data retrieval and manipulation

The European Bioinformatics Institute (EBI) provides Gene Ontology Association (GOA) files for completely sequenced proteomes. We downloaded the GOA files (http://www.ebi.ac.uk/GOA/proteomes, November 2009) for a total of 1,595 organisms spanning superkingdoms Archaea, Bacteria and Eukarya. We filtered out proteomes that were below the 50% coverage, with coverage defined as the number of proteins assigned to terminal GO_{TMF} terms divided by the total number of proteins in a GOA file. We also removed multiple occurrences of the same species, reducing the dataset to 638 non-redundant proteomes. To minimize sampling bias of proteomes between the three superkingdoms, we sampled only one bacterial species per genus, preferentially type strains. In the case of the other two superkingdoms, we retained all sampled proteomes without exclusion. For the remaining 358 proteomes, we studied organism lifestyles using various online resources (e.g. Genomes Online Database (GOLD) [121]), and published data [21,27,38]. Out of the total 358 organisms, 249 were identified as free-living and 109 either facultative parasitic or obligate parasitic. We generated two datasets: (1) *total* with the complete set of 358 proteomes, and (2) *free-living* with only 249 proteomes. We downloaded the OBO flat file from the GO database that gives the standard representation of gene ontologies (http://www.geneontology.org/GO.downloads.shtml; November, 2009). Out of the total 8,659 redundant MF terms that were defined in the OBO file, 1,708 were non-redundantly classified as parents and 3,396 as terminal nodes. We scanned for the presence of 3,396 terminal terms in both the *total* and *free-living* datasets. This resulted in 2,050 and 2,039 GO_{TMF} terms identified in the *total* and *free-living* datasets, respectively. Terms that were not present in the GOA files of our sampled proteomes were excluded from the analysis.

Phylogenomic analysis

For both the *total* and *free-living* datasets, we calculated a genomic census by counting the number of times each GO_{TMF} term was present in every functionome. We defined this count as the 'genomic abundance' value [20,28]. This value varies across functionomes as complex organisms encode extremely diverse and specialized functions in comparison to simple organisms. To account for the differences among functionome sizes and unequal variances, and

also because most phylogenetic software allow only up to 32 character states, we normalized the genomic abundance values in an alphanumeric format from 0 to 9 and A to V using the following formula [20,38]:

$$g_{ab\_norm} = \text{Round}\ [\ln(g_{ab}+1) / \ln(g_{max}+1) * 31\ ]$$

Using this formula, the genomic abundance value for each terminal GO_{TMF} term in every functionome (g_{ab}) is standardized by the maximum value in the matrix (g_{max}) and normalized to a scale from 0 to 31. The result is a matrix with rows representing functionome names (taxa) and columns representing GO_{TMF} terms (characters) with 32 possible character states (i.e. normalized abundance values) (Figure 2.1). These character states are linearly ordered, carry equal weight, and are compatible with the phylogenetic reconstruction software PAUP* ver. 4.0b10 [44]. Linear ordering of character states does not violate the assumption of character polarity as changes in both directions, forward (e.g. 18 to 24) and reverse (e.g. 22 to 9), are allowed and found to be frequent when traced on the branches of ToL (Chapter 1). These changes count towards tree length when maximum parsimony (MP) was used as the optimality criterion to search for the best possible tree with the minimum number of character state changes (Figure 2.1). MP is the most appropriate optimality criterion for analysis of this kind since we pool the entire set of known genes into a single study. These genes are evolving with different evolutionary rates and in such instances MP is shown to give better performance than any other tree reconstruction method [101]. Furthermore, convergence is less likely when using large number of multistate characters [101,122]. Trees were polarized using the ANCSTATES command in PAUP* and 0 was specified as the ancestral character state. We assumed that ancient functionomes encoded only a handful of functions and progressively enriched their repertoires along the evolutionary timeline [88]. Trees were rooted using the Lundberg method [47] that places the root at the most parsimonious location without the need to specify the outgroup taxa (see [27] for methodological explanations).

The phylogenetic error (i.e. effect of non-vertical evolutionary processes such as HGT and/or convergent evolution) was estimated by calculating retention indexes (r_i) for individual GO_{TMF} terms using the 'DIAG' option in PAUP*. The r_i indicates fit of characters to the phylogeny and is evaluated on a scale from 0-1 [123]. Higher r_i values indicate better fit of phylogenetic characters and thus lower probability of non-vertical inheritance. The statistical

significance of differences between two distributions of r_i values was evaluated by the Student's unpaired two-tailed t-test. The reliability of the phylogenetic trees was evaluated by 1,000 non-parametric bootstrap (BS) replicates.

To measure the degree of monophyly of individual taxonomic groups on a phylogenetic tree, we calculated the genealogical sorting index (GSI) using the module *GenealogicalSorting* ver. 0.92 of the R package ver. 2.15.1 with 10,000 permutated replicates [124]. The maximum GSI value of 1 signals the complete monophyly of a given taxonomic group and values close to zero indicate increase of dispersal. Trees were visualized using Dendroscope ver. 3 [49].

Reconstruction of rRNA trees

We downloaded the manually curated aligned sequences of rRNA genes (16S for Bacteria and Archaea, and 18S for Eukarya) for 231 out of 249 genomes of the *free-living* dataset from the SILVA database, release 111, which are reliably curated by considering alignment quality and phylogenetic relationships [125]. For the remaining 18 genomes, reliable alignments could not be extracted due to differences in naming conventions. All of the 231 rRNA sequences in the alignment were nearly complete in length (longer than 1,200 bp). The alignment file was imported into the ModelTest program [126] to determine the most appropriate nucleotide substitution model. Based on the results corresponding to the hierarchical likelihood ratio test, *GTR+I+G* was identified as the candidate model accounting for both the proportion of invariant sites and gamma-distributed rate variation [127]. Sequence alignment and model parameters were then imported into PAUP* to reconstruct a Neighbor-Joining (NJ) tree [128]. For individual phyla of the NJ tree, GSI values were calculated and compared with the MP trees.

Reconstruction of network trees

Network diagrams are useful indicators of any conflicts that may be present in the phylogenetic model and the reconstructed trees [129]. These networks are also termed neighbor-nets or network trees. We generated phylogenomic networks using the Neighbor-Net algorithm implemented in the SplitsTree package ver. 4.12.6 [130]. We transformed the abundance matrices (described above) into occurrence (i.e. presence/absence) matrices for calculation of distance-based phylogenies. To evaluate the amount of 'vertical' phylogenetic signal present in our data we calculated the delta (δ) score, a measure of the reticulation levels of networks on a scale from 0 to 1 [131]. A δ-score of 0 indicates a fully bifurcating tree while a value close to 1

means complete absence of vertical phylogenetic signal or a full network [131]. Example of modern use of neighbor-nets and δ-scores can be found in recent evolutionary studies of language [132] and culture [133].

Enrichment test for HGT

To quantify the degree of HGT affecting phylogenetic trees, we compared 249 *free-living* proteomes to the prokaryotic proteomes listed in the horizontal gene transfer database (HGT-DB; [134]). Only 72 out of 249 proteomes were cross-listed along with GenBank identifiers (GIs) for potential horizontally transferred proteins (HTPs). These proteins were however listed with their UniProtKB IDs in the corresponding GOA files. We therefore converted the GIs of HTPs to UniProtKB IDs using the online ID MAPPING tool of UniProt (http://www.ncbi.nlm.nih.gov/Taxonomy/TaxIdentifier/tax_identifier.cgi) and determined GO associations for HTPs present in these 72 proteomes. To evaluate whether a GO_{TMF} term should be regarded as a horizontally transferred character or not, we examined how many HTPs corresponded to proteins of a GO_{TMF} term. The degree of the association between HTPs and GO_{TMF} terms was estimated by conducting a statistical test using the hypergeometric distribution [38,135]. We compared the total number of HTPs that were associated with a GO_{TMF} term (sample) to the total number of proteins present in the 72 proteomes regardless of HGT (background). The statistical significance of enrichment was evaluated at the 95% confidence level and using the following equation,

$$P(X = k) = \frac{\binom{M}{k}\binom{N-M}{n-k}}{\binom{N}{n}}, \text{ where } \binom{a}{b} = \frac{a!}{b!(a-b)!}$$

where, k indicates the multiple occurrences of a GO_{TMF} term associated with HTPs in the sample; n indicates the total redundant numbers of all GO_{TMF} terms in the sample; M indicates the multiple occurrences of a GO_{TMF} term associated with HTPs in the background; N indicates the total redundant numbers of all GO_{TMF} terms in the background; and P ($X = k$) indicates the probability of enrichment.

Results

ToL reconstructions describe the evolution of functionomes

Since the physiology of an organism is defined by the biological functions of its molecular components, a tree with functionomes as taxa is the closest that is possible to a bonafide tree of organisms and a bonafide ToL. We therefore reconstructed rooted ToLs from genomic abundance counts of terminal GO_{TMF} terms in the *total* (Figure B1) and *free-living* (Figure B2) datasets, which were used as phylogenetic characters. The *total* dataset included a repertoire of 2,050 GO_{TMF} terms from 358 organisms encompassing 47 Archaea, 288 Bacteria, and 23 Eukarya with both free-living and non-free-living (i.e. parasitic/symbiotic) lifestyles. ToLs reconstructed from the *total* dataset highlighted the bias associated with the inclusion of the functionomes of organisms that were not free-living (and interact with hosts) as most of these taxa occupied the most basal positions in the tree (red squares in Figure B1). These taxa included notable parasites such as *Nanoarchaeum equitans* (archaeal parasite), *Guillardia theta*, (nucleomorph with a highly reduced genome; Cryptophyta, marine plankton, Eukarya), *Candidatus* Carsonella rudii (gamma-proteobacteria) and *Candidatus* Sulcia muelleri (beta-proteobacteria) (both symbionts of sap-feeding insects) [21]. In addition, functionomes from a number of bacterial parasites were clustered at the base of the bacterial group including Actinobacteria, Bacteroidetes, Chlamydia, Firmicutes, Fusobacteria, Spirochetes, and various proteobacterial symbionts (Figure B1). Moreover, functionomes from Eukarya did not form a distinct superkingdom but rather appeared as a subgroup derived from Bacteria, highlighting important inaccuracies in the topology of this tree.

The link between parasitism and genome reduction has been explained previously [20,21,27] and is known to bias tree reconstructions and affect the topology of ToLs [38]. For example, organisms that engage in obligate parasitism can loose nearly all of their metabolic genes and depend upon the host for survival [21]. These idiosyncratic host-mediated tendencies of genome reduction (scattered in parasitic lineages throughout the ToL) affect the functional makeup of proteomes and complicate phylogenetic reconstruction. They also add a bias to our evolutionary model, which based on the principle of continuity assumes that ancestral functionomes had a simpler repertoire of molecular functions that progressively became richer. Because parasitic/symbiotic organisms harbor highly reduced genomes, our model favored their

placement at basal positions of the tree. To avoid these biases, we examined the lifestyles of the 358 organisms of the *total* dataset and excluded 109 organisms with parasitic/symbiotic lifestyles. The remaining 249 organisms harbored a functional repertoire of 2,039 GO_{TMF} terms. This *free-living* dataset included functionomes from 45 Archaea, 183 Bacteria and 21 Eukarya.

ToLs reconstructed from the *free-living* dataset, now free from the effects of problematic taxa, supported the division of living organisms into three distinct superkingdoms: Archaea, Bacteria and Eukarya (Figure B2). Archaeal lineages rooted the tree paraphyletically and made up the most ancient superkingdom (read below). In turn, Bacteria and Eukarya formed monophyletic groups that shared a common ancestor separated from Archaea by 89% BS (Figure B2). We note that BS values depend on number of taxa and are generally expected to be low in ToLs of these sizes. This fact should be taken in consideration when evaluating the significance of phylogenetic relationships. We also note that genome reduction is not restricted to only parasitic and symbiotic organisms. Gene loss may also occur in free-living cells, albeit at lower levels. Robustness of our phylogenetic methodology against these cases is supported by the phylogenetic positions of *Pelagibacter ubique* (marine alpha-proteobacteria) and *Prochlorococcus marinus* (cyanobacteria), both well-documented examples of genome reduction in free-living organisms [136,137]. Previous phylogenetic studies based on gene sequences showed that *P. ubique* and *P. marinus* were closest to *Zymomonas mobilis* and *Synechococcus* sp., respectively. Unlike *P. ubique* and *P. marinus* (genome sizes ca. 1.3 and 1.7 Mbp, respectively), *Z. mobilis* and *Synechococcus* sp. have larger genomes (ca. 2 Mbp and 2.5 Mbp, respectively) and are free from genome reduction. Nevertheless, *P. ubique* and *P. marinus* are closest to *Z. mobilis* and *Synechococcus* sp., respectively, in the ToLs reconstructed from both the *free-living* (Figure B2) and the *non-HGT* datasets (Figure B2; read below). This strongly supports the claim that ToLs reconstructed by genomic abundance are robust against inclusion of reduced free-living functionomes. In fact, genome reduction in free-living organisms is mostly limited to auxiliary genes, still allowing most of essential genes to encode core molecular functions. Since functionally important genes largely represent the genomic abundance of a functionome, genome reduction of free-living organisms may result in a small decrease of their genomic abundance. Consequently, ToLs reconstructed by genomic abundance would only be marginally affected by the inclusion of reduced free-living functionomes.

Identification of GO_{TMF} terms associated with horizontally transferred proteins

To better resolve phylogenomic relationships, problematic characters that are acquired via HGT and contribute to homoplasy must be also excluded [27,38]. HGT is believed to have played an important role in microbial evolution, especially in Bacteria [25]. Because the *free-living* dataset included a large number of bacterial functionomes (73%), ToLs built from this set could also lead to confounding results. Horizontally transferred proteins (HTPs) do not contribute to 'shared and derived' GO_{TMF} terms, which are the backbone of vertical phylogenetic signatures, and can only add phylogenetic noise. Their exclusion is thus justified at the expense of reducing phylogenetic accuracy. To define GO_{TMF} terms that were significantly associated with HTPs, we evaluated the enrichment of HTPs for individual GO_{TMF} terms using the hypergeometric distribution, which was already successfully applied to evolutionary studies of this kind [38,135]. We identified HTPs in 72 out of 249 free-living organisms that were cross-listed in the HGT-DB [134] and extracted their GO associations. We then compared the enrichment of these GO_{TMF} terms (sample) to the enrichment of the rest of the GO_{TMF} terms in the 72 functionomes (background) and evaluated statistical significance at 95% confidence level (see Methods). A total of 115 out of the 2,039 GO_{TMF} terms were significantly associated with HTPs ($P < 0.05$). Exclusion of these terms from the *free-living* dataset resulted in 1,924 phylogenetic characters. This new *non-HGT* dataset was used to reconstruct a ToL that described the evolution of functionomes from 249 free-living organisms and minimized the effect of HGT (Figure 2.2). The new tree was mostly congruent to the tree reconstructed from the *free-living* dataset (Figure B2) (read below).

Phylogenomic patterns

The optimized ToL generated from the *non-HGT* dataset supported the tripartite nature of the living world and monophyletic Bacteria and Eukarya, which were grouped as sister taxa (61% BS) emerging from paraphyletic Archaea (Figure 2.2). The ToL also uncovered remarkable phylogenomic patterns:

(i) A hyperthermophilic origin of diversified life in Archaea. A closer examination of the basal archaeal lineages of the ToL with splits exhibiting 50-90% BS showed that they embodied organisms belonging to the orders Desulfurococcales and Thermoproteales of Crenarchaeota. They included *Desulfurococcus kamchatkensis*, *Hyperthermus butylicus*, *Staphylothermus marinus* and *Thermofilum pendens*. *Desulfurococcus* is a genus of thermophilic, organotrophic

and anaerobic archaea generally found in hyperthermic habitats such as deep-sea thermal vents and subterranean hot springs [138]. *T. pendens* is a thermophilic and moderate acidophile archaeon isolated from a solfataric hot spring that uses sulfur and peptides as energy source [139]. *S. marinus* and *H. butylicus* are also hyperthermophile archaea belonging to the Desulfurococcales that can be sulfur reducing and live in deep-sea vents and hot solfataric floor habitats [140,141]. While the hyperthermophilic origin of diversified life has always been associated to the rise of Bacteria, our finding that the root of the ToL lies in hyperthermophilic Archaea is very significant.

(ii) Cohesive archaeal orders but non-cohesive major archaeal groups. Organisms in well-recognized archaeal orders were unified but with widely ranging supports, from well supported clades in Halobacteria (100% BS), Sulfolobales (98% BS), Thermococci (89% BS), Methanococci (82% BS), to moderate support for the branch grouping of both Methanomicrobia and Methanobacteria (74% BS) and Thermoplasmata (64% BS). However, support for deeper branches unifying these orders was consistently low. We found that crenarchaeal organisms belonging to the order Sulfolobales were derived and appeared associated with Thaumarchaeota, while the rest of archaeons belonging to Euryarchaeota occupied intermediate basal positions in the tree, together with Korarchaeota.

(iii) A non-thermophilic origin of Bacteria. Groupings of phyla in the ToL favored the non-thermophilic origin of the bacterial superkingdom. The tree placed the anaerobic rod-shaped Bacteroidetes and some members of the PVC superphylum such as Verrucomicrobia in the most basal positions, linked to a more derived actinobacterial phylum. Some well-recognized bacterial phyla were strongly unified with good to moderate support, including the Chlorobi (100% BS), Synergistetes (100% BS), Chloroflexi (99% BS), epsilon-proteobacteria (93% BS), Cyanobacteria (69% BS) and Aquificae (57% BS), while other were unified with poor bootstrap supports (<50% BS), including Thermotogae, delta-proteobacteria and a large group of gamma-proteobacteria. Firmicutes appeared in more basal positions than alpha-proteobacteria, beta-proteobacteria, and gamma-proteobacteria, none of which formed cohesive groups. The thermophilic Aquificae and Thermotogae were quite derived when compared to organisms of the basal PVC group.

(iv) Monophyletic relationships in major eukaryal groups and close relationship between Plants and Metazoa. Eukarya formed a strong monophyletic group (100% BS; Figure 2.2). Metazoa, Plants and Fungi were also monophyletic with taxa in the individual groups well positioned. Remarkably, the ToL of functionomes recovered a sister taxa relationship of Metazoa and Plants (59% BS). At the time of the analysis, the functionomes of only two flowering plants (*Arabidopsis thaliana* and *Vitis vinifera*) with coverage of more than 50% were available. While the sister relationship between Metazoa and Plants may be due to limited sampling of taxa, the close relationship between the two groups was also recovered in previous evolutionary studies that focused on abundance of protein domains [20,43]. Recently, a ToL reconstructed from the abundance of conserved protein domains in 420 free-living organisms also identified a close relationship between Metazoa and Plants [38]. In this study, authors sampled a large number of eukaryal proteomes including 64 Metazoa, 44 Fungi, 16 Protista, and 9 Plants. The ToL revealed that Fungi was distant from both Metazoa and Plants, while the latter two were clustered in close proximity and separated by 5 animal-like protist proteomes. This suggests that inclusion of more eukaryal functionomes, especially of protists, can change existing deep phylogenetic relationships in Eukarya. However, the topological consistency between the functionome-based and the protein domain-based ToLs at least supports that Plants is a closer evolutionary relative of Metazoa than Fungi. It is therefore likely that plants and animals share physiological similarities and encode a functional apparatus that is quite similar. It would be interesting to validate this hypothesis in future studies.

Evaluating the degree of monophyly in phylogenetic trees

To quantify and compare the historical relationships among groups of organisms in the ToLs reconstructed from the *free-living* and *non-HGT* datasets, we calculated the degree of monophyly (GSI values) for individual groups consisting of at least five functionomes (Table 2.1). Six out of 14 groups of the ToL reconstructed from the *non-HGT* dataset (including Crenarchaeota, Actinobacteria, and all proteobacterial phyla) had larger GSI values than the one reconstructed from the *free-living* dataset. In turn, only two phyla (Euryarchaeota and Bacteroidetes) exhibited larger GSI values in the ToL reconstructed from the *free-living* dataset. In case of the remaining six groups (Chlorobi, Cyanobacteria, Firmicutes, Thermotogae, Fungi, and Metazoa), both trees showed the same degree of monophyly. Since HGT occurrences in proteobacterial genomes are very common [142], increased GSI values of proteobacterial phyla

in the ToL derived from the *non-HGT* dataset indicated that the exclusion of HTPs characters increased significantly the accuracy of phylogenetic statements despite of reducing cladistic information.

Because rRNA genes are highly conserved and commonly used in sequence-based phylogenies, we also compared the degree of monophyly of the *non-HGT* tree to the NJ tree reconstructed from 16S and 18S rRNA gene sequence alignment. A comparison of GSI values revealed that groups in the *non-HGT* tree were generally better supported (Table 2.1). Overall, seven out of 14 groups had higher GSI values in the *non-HGT* tree compared to the rRNA tree including, Crenarchaeota (0.80 vs. 0.60), Actinobacteria (0.88 vs. 0.87), Bacteroidetes (0.83 vs.0.28), Firmicutes (0.82 vs. 0.72), gamma-proteobacteria (0.74 vs. 0.41), Fungi (1.00 vs. 0.21) and Metazoa (1.00 vs. 0.56). These included both the very basal (e.g. Crenarchaeota) and derived (e.g. Fungi and Metazoa) branches of the ToL. In contrast, rRNA tree performed poorly in resolving the very derived branches of Fungi (GSI = 0.21) and Metazoa (GSI = 0.56). Five out of 14 groups had higher GSI values in the rRNA tree and included proteobacterial phyla (alpha-proteobacteria [0.70 vs. 0.66], beta-proteobacteria [0.87 vs. 0.48], delta-proteobacteria [1.00 vs. 0.51]), Thermotogae (1.00 vs. 0.80) and Euryarchaeota (0.92 vs. 0.70). Chlorobi and Cyanobacteria had GSI value of 1.00 in both trees.

This exercise revealed that the *non-HGT* tree performed superior to the rest of the reconstructed trees and that the use of GO definitions as phylogenetic characters served better in resolving monophyletic relationships. We argue that trees built from the entire functional toolkit (e.g. *free-living*, *non-HGT*) are more powerful in charting organismal relationships than those built from limited character sets (e.g. *info* tree; read below) or a single molecule (rRNA tree), especially when considering that the entire functional apparatus of an organism approximates the physiology of that organism and truly depicts a ToL. In contrast, rRNA represents only one of the three classes of rRNA molecules that make structural components of ribosomes and does not represent the entire evolutionary history of an organism (). Therefore, inferences regarding entire systems (i.e. organisms) should include all the individual components that make up that system (i.e. protein domains, functional definitions) rather than only a single (albeit ancient and central) molecule. Hence, from hereinafter, we will only focus on elaborating phylogenies resulting from the census of molecular functions as they allow to make systemic comparisons among organisms and enable the evolutionary study the of organisms as biological systems.

Exclusion of problematic taxa and horizontally acquired characters improved phylogenetic reconstructions

We inspected the reliability of phylogenetic trees recovered from the census of molecular functions by selecting only 120 GO_{TMF} terms that were involved in informational processes, including transcription and translation. This character set was used to build a new ToL. It has been proposed that information-related genes are refractory to the effects of HGT [25]. A single most parsimonious tree reconstructed from the limited set of informational GO_{TMF} terms in the 249 free-living functionomes was largely congruent with the ToL reconstructed from the *non-HGT* dataset (Figure B3). This tree also favored the groupings of organisms into three superkingdoms and was rooted paraphyletically in Archaea. While Korarchaeota clustered with the eukaryal clade, the tree fared well in terms of overall groupings among phyla (Figure B4). The number of monophyletic phyla recovered was however lower than the number recovered in the *non-HGT* ToL (Figure B4). Furthermore, only three groups of the tree of information processes (i.e. Euryarchaeota, alpha-proteobacteria, beta-proteobacteria) had larger GSI values than the *non-HGT* ToL (Table 2.1). One explanation is the lesser number of phylogenetic characters used to reconstruct the tree (120 versus 1,924). In general, using large number of characters improves phylogenetic reconstruction [102,143]. To test this, we extracted 1,000 random samples each consisting of 120 GO_{TMF} terms from the 1,843 parsimoniously informative non-HGT characters and generated 1,000 trees. We noted that most of the random *non-HGT* ToLs still had more monophyletic phyla compared to the tree of information processes (data not shown). It is therefore desirable to generate trees from the entire functional toolkit and not just a specific functional repertoire, as explained above.

To further investigate the reliability of phylogenetic reconstructions, we compared the r_i distributions of ToLs recovered from functionomic data (Figure 2.3A). These included trees derived from the *total* (all taxa and characters included), *HGT* (only 115 HTP-derived GO terms included) and *non-HGT* (both the problematic taxa and characters excluded) datasets. In general, higher r_i values support better fit of phylogenetic characters to the phylogeny and thus lower probability of non-vertical inheritance. The boxplots indicated that the best trees were recovered using the *non-HGT* dataset (Figure 2.3A), supporting previous results. In contrast, *HGT* trees indicated the worst fit and were on average distributed with the lowest r_i values. A comparison between the *HGT* and *non-HGT* trees was statistically significant ($P < 0.05$) (as expected)

(Figure 2.3A) suggesting that any confounding effects resulting from HGT were controlled in the *non-HGT* trees.

Finally, we confirmed the validity of our MP-based ToLs and tested for any conflicts between our evolutionary model and phylogenomic trees by reconstructing phylogenomic networks. Our phylogenomic model assumes that functionomes became progressively more complex; i.e. we consider gene gain and loss, gene rearrangements and gene duplications to be the major evolutionary forces shaping the functionomes of living organisms [6,27,43]. When the phylogeny involves complex evolutionary processes, a more abstract network representation can be used to test any conflicts between the model and the tree [129]. Phylogenomic networks generated from the occurrence data (i.e. presence or absence of GO_{TMF} terms) for the *total*, *HGT* and *non-HGT* datasets validated the *non-HGT* dataset and highlighted important shortcomings of the *HGT* dataset (Figure 2.3B). Phylogenomic networks generated from the *total* dataset included archaeal and eukaryal parasites (*Nanoarchaeum equitans* and *Guillardia theta*) that were clustered within Bacteria clearly suggesting a revision of the evolutionary model. In contrast, the *non-HGT* network supported the three-superkingdom classification system with no contamination of taxa (Figure 2.3B). Finally, the *HGT* network constructed from the 115 HTP-derived GO terms failed to re-enact a tree-like structure with true bacterial and eukaryal groupings and showed that the HTP-derived GO terms did not complicate Archaeal relationships (Figure 2.3B). This was a significant result and raised important questions. First, it questions the existence of pervasive HGT within and between microbes. Second, it shows that the exclusion of HGT-derived GO terms significantly improved the phylogenies of *non-HGT* dataset. Third, it challenges the existence of fundamental organismal fusions used to explain evolutionary reticulation. All of these observations are significant and mandate future investigation.

To test if the poor resolution of the *HGT* network was not due to the limited number of phylogenetic characters used for its reconstruction, we randomly sampled 115 GO_{TMF} characters from the *non-HGT* dataset and prepared 1,000 random files for network analysis. We discovered that the majority of the random networks partitioned the organisms into three unified groups and did not suffer from limited sampling (Figure B5). Thus in light of our results, the poor resolution of the *HGT* network should be considered significant. To identify taxa that were contributing to reticulation patterns in the networks, we calculated δ-scores for individual phyla and

superkingdoms. The δ-distribution is shown for the *non-HGT* network that revealed interesting but expected patterns (Figure 2.3C). Both microbial superkingdoms were distributed with high δ-values with scores ranging from 0.27-0.35 in Archaea and 0.30-0.42 in Bacteria. In contrast, the contribution to genetic exchange of eukaryal functionomes appeared minimal (0.16-0.34) (Figure 2.3C). All the comparisons were statistically significant at 95% confidence level and suggested that the rates of non-vertical evolutionary processes or HGT varied significantly between superkingdoms. The degree of reticulation in superkingdoms increased in the order Eukarya, Archaea and Bacteria (Figure 2.3C), suggesting a similar trend for the HGT correlate. The lowest δ-score averages were observed in mammals and primates (δ = 0.16-0.17) in Eukarya, Methanococci, Methanobacteria and Thermococci (δ = 0.28-0.29) in Archaea, and Thermotogae and Dictyoglomi (δ = 0.31-0.32) in Bacteria. A comparison of δ-scores for the different bacterial groups confirmed that the majority of the major bacterial taxonomic groups (e.g. Gemmatimonadetes, Verrumicrobia, Bacteroidetes, Acidobacteria, and others; Table 2.2) were the largest contributors to genetic exchange. In contrast, eukaryal superkingdoms appeared to be best supported in the ToLs with lowest δ-scores. Finally, archaeal phyla were supported with intermediate values (Table 2.2). The overall δ-score for the *non-HGT* network was 0.33, in comparison to 0.34 for the *total* network and 0.39 for the *HGT* network, clearly identifying *non-HGT* networks and trees to be best resolved.

These experiments revealed that the ToL derived from the *non-HGT* dataset reflected phylogenomic relationships most accurately. This dataset is free from the effects of parasitic organisms and is minimally affected by non-vertical evolutionary processes. We conclude by mentioning that our phylogenomic approach is robust against unequal sampling of proteomes per superkingdoms that can lead to incorrect parsimonious trees due to long-branch-attraction [38]. Therefore, the relatively large number of bacterial proteomes in the *non-HGT* dataset (once the HGT-derived characters are excluded) is not expected to bias phylogenomic relationships, as reported previously [38].

GO coverage does not bias phylogenetic relationships

In this study, we included only organisms with functionomes that provided at least 50% coverage of molecular functions. We note that many of the sampled functionomes were annotated in reference to the experimentally verified GO annotations in few model organisms.

Thus, large GO coverage differences in functionomes could reflect the similarity of functionomes to model organisms and thus bias the phylogenetic relationships. However, the functionomes we sampled had a mean GO coverage of 59.23% and a standard deviation of 5% (Figure B6). The small variance indicated that the distribution of GO coverage was quite even across functionomes. Furthermore, the coverage of most model organisms (e.g. *Homo sapiens* of 62%, *Arabidopsis thaliana* of 51%, *Mus musculus* of 67%, *Drosophila melanogaster* of 65%, etc.) was quite similar to the mean and within the upper and lower whiskers of the GO distribution. There were only few outliers: *Saccharomyces cerevisiae* (82% in GO coverage), *Rattus norvegicus* (78%) and *Gallus gallus* (78%) (Figure B6). These results indicate that the degree of GO annotation for non-model organisms is comparable to that for model organisms. In other words, the GO coverage of the functionomes we sampled shows that electronic GO annotations (mostly for non-model organisms) are quite saturated in comparison to experiment-based annotations (most for model organisms).

Although the GO coverage of most functionomes was close to the mean coverage, we observed that few taxonomic groups were associated with relatively large variance of coverage across the three superkingdoms. Remarkably, the functionomes of these taxonomic groups were still grouped together in the *non-HGT* tree (Figure 2.2). For example, three *Pyrococcus* functionomes (i.e. *P. abyssi*, *P. furiosus*, and *P. kodakaraensis*) that had 63%, 58% and 52% coverage, respectively, were clustered monophyletically as a single genus (Figure 2.2). A more extreme case in Eukarya was the monophyletic clade of *S. cerevisiae* and *Pichia stipitis* that belongs to Saccharomycetaceae. Although the GO coverage of the two species was significantly different (82% for *S. cerevisiae* and 55% for *P. stipitis*), they were still clustered together in the *non-HGT* tree (Figure 2.2). In Bacteria, previous phylogenetic studies have supported the strong monophyly of Cyanobacteria. Remarkably, all six cyanobacterial functionomes with GO coverage ranging from 51 to 57% grouped together. Based on the evidence from balanced distributions of GO coverage and phylogenetic groupings of closely related taxa with large variance GO coverage, we conclude that the extent of GO annotations did not significantly affect positioning of organisms in the ToL. Instead, our previous phylogenetic experiments showed that tree topologies of molecular functions largely depend on how differently individual GO terms are assigned to a functionome but not on how many GO annotations are assigned to a

functionome [88]. This implies that GO coverage had minimal effect on phylogenetic placements.

Interplay between genomic abundance and occurrence

Different evolutionary forces are responsible for the accumulation of functions in genomic repertoires [20,43], including gene duplications, gene rearrangements and HGT. These events lead to a direct increase in the genomic abundance of genes and corresponding molecular functions [27]. Abundance is therefore a naturally occurring biological process that is valuable for reconstructing phylogenies [5,37]. In contrast, occurrence-based approaches involve non-redundant representations of genes (and their functions) that generally result in more balanced topologies [144]. We observed that both abundance and occurrence of GO_{TMF} terms were correlated and resulted in congruent classifications (Figure 2.4). For instance, plotting occurrence and abundance of GO_{TMF} terms against their distribution in proteomes (distribution index or *f*-value = number of functionomes encoding a GO_{TMF} term / total number of functionomes) revealed interesting relationships (Figure 2.4A).

The majority of the GO_{TMF} terms (~1,300 or >60%) were not conserved across taxa ($f < 0.1$) but were distributed with low abundance values (~200/functionome). These terms represent molecular functions that are relatively new to the functional toolkits of proteomes and are not universally distributed. They also correspond to organism-specific functions that have been acquired late in evolution. In contrast, GO_{TMF} terms that were universally present ($0.9 < f < 1.0$) were very few in number (~185) but had the highest abundance values (~25,000/functionome) (Figure 2.4A). These terms represent ancient molecular functions that are vital for cellular life and are conserved across most taxa (e.g. ATPase activity, helicase activity). Excluding the two extremes (i.e. most recent and most ancient) resulted in both abundance and occurrence being evenly distributed and showed there was no bias favoring one or the other. This analysis supported our choice to study the terminal terms that provided very high resolution for differentiation of organismal relationships.

When plotted individually for each functionome, we found a strong correlation between the two concepts (Figure 2.4B). Organisms followed a trend from simplicity towards complexity in biological organization, beginning with the simplest functionomes of Archaea, closely followed by a diverse range of bacterial and eukaryal functionomes (Figure 2.4B) and ending

with the extraordinarily rich functionomes of *Homo sapiens*, *Mus musculus*, *Arabidopsis thaliana*, *Rattus norvegicus*, *Danio rerio*, *Bos taurus*, and *Drosophila melanogaster*, which appeared as outliers but were distinguished by both abundance and occurrence parameters. This result is in line with a similar analysis of protein domain abundance and occurrence [20,27]. We conclude that genomic abundance and occurrence are positively correlated and that using abundance enhances deep phylogenetic signal [37] in the study of molecular functions.

Discussion

A new ToL with taxa that better-depicts the physiology of organisms

Using an atypical application to a well-established cladistic methodology, here we reconstructed rooted ToLs without the use of outgroups directly from a genomic census of biological functions. These trees are unprecedented. They describe the evolution of entire repertoires of molecular functions and have an evolutionary arrow built into their driving evolutionary model. This is highly significant. Thus far, ToLs are extrapolations of molecular trees that rest on the assumption that the essence of an organism can be appropriately depicted by a single molecule or a repertoire of molecules that are hopefully minimally affected by HGT [111]. In particular, the small subunit of rRNA has been used as gold standard despite of representing only one of three RNA subunits that typically, and together with dozens of ribosomal proteins, make up the ribosomal ensemble. The finding that rRNA coevolves with ribosomal proteins and that the ribosome is younger than tRNA and important enzymes (e.g. aminoacyl-tRNA synthetases) and has a protracted history [97] complicates the arguments of the evolutionary centrality of one or a set of its components and the functional link between the ribosome and the organism. Instead, the functionome, suitably defined by ontological terms, approaches the entire collection of functions of an organism and is therefore unbiased by preconceptions on molecular biology and biochemistry. The abundance-based approach also shields deep phylogenomic relationships of functionomes from the effect of HGT and functional recruitment [88]. Functions that are laterally transferred or are recruited must be first fixed and then amplified to high levels in genomic evolution if they are to have an impact on the deep branches of the ToLs. In other words, HGT or recruitment of functions that are abundant and are ancient will have little impact on the basal topologies of the trees. In contrast, small changes in genomic abundance of functions that are rare, of recent ancestry and specific to selected lineages can only significantly affect very derived branches of the trees. This and other properties of the new reconstruction method makes trees of functionomes excellent complements to trees of molecules derived from sequence analysis, which perform best when comparing closely related organisms.

We confirmed the validity of our phylogenomic statements by comparing the degree of monophyly with the canonical reference tree, building distance-based phylogenomic networks,

excluding problematic taxa and HTP-linked characters, and evaluating phylogenetic reticulation due to non-vertical evolutionary processes such as HGT, endosymbiosis and recruitment. Remarkably, we observed cohesiveness and robustness of Archaeal relationships in phylogenomic networks that question the idea that HGT between microbes (e.g. between Archaea and Bacteria) occurs at dramatically high levels [10,145] and challenges the fusion model for the origin of eukaryotes that attributes the origin of Eukarya to a primordial fusion event between archaeal and bacterial cells (see [146] and references therein). For example, reticulation measures in networks (δ-score) showed minimal reticulation in Eukarya, intermediate levels in Archaea, and as expected, maximal reticulation impact in Bacteria (Figure 2.3C). However, reticulation levels of some euryarchaeal (e.g. Methanococci and Methanobacteria) and crenarchaeal (e.g. Sulfolobales and Thermoproteales) orders in Archaea were not so far away from reticulation levels in plants, and reticulation of several bacterial orders such as Thermotogae, Firmicutes and Chlorobi were comparable to average levels of archaeal reticulation (e.g. in crenarchaeal orders) (Table 2.2). In particular, gamma-proteobacteria harbor species that exhibit unprecedented HGT levels, such as *Shewanella baltica*, which exchanges up to 20% of their entire core and auxiliary genome in short time frames [147]. These processes of rapid adaptation through massive acquisition of genes, which are common in the ocean [148] and in other aquatic environments [149], are not reflected in the δ-scores of the *Shewanella* genus (e.g. *S. putrefaciens;* δ = 0.34) or the gamma-proteobacterial order (δ = 0.36), which are comparable to those of *Arabidopsis thaliana* (δ = 0.32), *Saccharomyces cerevisiae* (δ = 0.34) and other eukaryotes. All of these results challenge the perception that reticulation and its HGT correlate is rampant in the long-term evolution of microbes.

The early thermophilic origin of Archaea

ToLs generated from the genomic census of molecular functions supported the view that Archaea was the first cellular superkingdom to appear in evolution (Figures 2.2 and B2). The archaeal rooting of the ToL has been recovered previously in a number of studies where the focus was on building reliable phylogenies using conserved structural information in protein and nucleic acid molecules [20,27]. ToLs built from proteomic abundance of domain structure and organization defined at different levels of structural conservation of the Structural Classification of Proteins (SCOP) [34] and CATH [150] classifications consistently displayed a paraphyletic

rooting in Archaea [6,14,20,38,75]. Similar results were obtained when building trees of RNA molecules from nucleic acid structure in 5S rRNA [151] and RNase P RNA [152] and from nucleic acid sequence and structure in tRNA [89]. More importantly, timelines of accretion of helical RNA substructures of tRNA [106] and 5S rRNA [151] uncovered two accretion pathways, one specific to Archaea and the other common to Bacteria and Eukarya.

Timelines of accretion in RNase P RNA showed that the most ancient substructures were universal and harbored the core catalytic activities of the endonuclease [152]. However, the first RNase P RNA substructures that were lost were specific to Archaea and this episode occurred before molecules were accessorized with superkingdom-specific substructures [152]. Evolutionary timelines of protein domain appearance in the protein world also showed the early loss of domains in Archaea prior to the appearance of superkingdom-specific domain structures in the analysis of domain and domain interactome evolution [6,14,20,38]. In fact, a phylogenetic tree reconstructed using 1,924 GO_{TMF} terms as taxa and 249 functionomes from free-living organisms as characters (*non-HGT* dataset) identified both the very ancient and derived GO_{TMF} terms [153]. In this study, most of the very ancient GO_{TMF} terms were only detected in the bacterial and eukaryal functionomes, but were completely absent in Archaea. While it can be argued that loss of ancient GO_{TMF} terms in Archaea could be a very recent event, the scenario does not seem very likely. This is because a single molecular activity is a product of multiple genes that have accumulated over the course of evolution. These genes multiply and increase their abundance in cells with the progression of time. Thus, loosing an *ancient* molecular function *late* in evolution is much more costly than loosing it earlier in evolution when genes have low abundance levels. In light of these considerations, we propose that genome reduction in thermophilic archaeal species was likely an ancient event that started very early in evolution and before the divergence of Bacteria and Eukarya. In comparison, the alternative scenario is not well supported by the distribution of conserved protein structures [20,27] and molecular functions (Chapter 3) in the proteomes and functionomes of contemporary organisms and is therefore less likely. Moreover, the paraphyletic archaeal root of the tree of life has also been suggested by early studies of interparalog distances of tRNA paralogs (aloacceptors) and paralogous pairs of aminoacyl-tRNA synthetases, which depend on intraspecies comparisons and are therefore intrinsic to each species [83,84]. These findings were further supported by additional polyphasic evidence [85,154].

The paraphyletic rooting of the ToL in Archaea is in striking disagreement with for example the canonical rooting in Bacteria that is achieved by the use of protein paralogs as mutual outgroups for central proteins such as aminoayl-tRNA synthetases, elongation factors (e.g. EF-Tu/EFG), ATPases, carbamoyl phosphate synthetases, and signal recognition particle proteins (reviewed in [155]). These paralogous rootings however are considered weak because of a number of problems and artifacts of sequence analysis (e.g. long branch attraction, mutational saturation, taxon sampling, HGT, hidden paralogy, historical segmental gene heterogeneity) and because they depend on the history of a small set of molecules out of the entire molecular repertoire of the cell. Distance-based approaches have also been used to build universal network trees from gene families defined by reciprocal best BLAST hits, which showed a midpoint rooting of the ToL between Bacteria and Archaea [156]. However, this rooting involves a complex optimization of path lengths in the split networks and critically assumes that lineages evolve at roughly similar rates. This diminishes the confidence of rootings of this kind, especially when considering the uncertainties of distances inferred from BLAST analyses and the fact that domains in genes hold different histories and rates of change. In fact, current approaches to rooting of molecular sequences bring almost insurmountable complexities that require novel conceptual frameworks, such as critical analysis of major evolutionary transitions (e.g. 'transition analysis'; [157]) to establish polarity of change [158] or the analysis of genomic insertions and deletions that are rare in paralogous gene sets [159]. However and as we commented above, the use of molecular sequence is problematic on many grounds, especially mutational saturation, violation of character independence by the mere existence of atomic structure, and different historical signatures in domains of multidomain proteins [37]. Similarly, establishing the validity of evolutionary transitions in polarization schemes can also be problematic and requires well-grounded assumptions for each transition that is used [158]. Remarkably, the assumptions of the intrinsic rooting scheme of molecular functions that we here present are supported by timely successions of major evolutionary transitions that increased biological complexity [160] and information transmission [161] when these transitions are mapped along a timeline of molecular functions (see Figure 4 in [88]).

The rooting of the tree of cellular life in Archaea is paraphyletic and requires explanation. While paraphyly could result from loss of phylogenetic signal or from primordial homoplasy-generating processes operating during the early differentiation of superkingdoms, trees are

particularly well supported at their base and the paraphyletic rooting is congruently obtained in different studies employing a diverse set of phylogenetic characters, from ontological terms to tRNA molecules. Thus, a more plausible explanation is that the early diversification of LUCA involved spatial colonization of unchartered environments that were ecologically unique to the individual primordial lineages [151]. This colonization was followed by selective reductive loss of genomic components [20] as the emerging archaeal lineages adapted to the different (initially auxinic) ocean and land environments of the late Archaean. This divergence-by-isolation scenario explains patterns of loss and gain of molecular structures and their associated functions in evolutionary timelines (e.g. [27,88]), which are for example responsible for delimiting the three evolutionary epochs proposed by Wang et al (2007) [20]: (i) an early architectural diversification epoch in which ancient molecules and their functions emerged and accumulated in proteomes as cells of a communal global ancestor became modularized into individual entities, (ii) a superkingdom specification epoch in which many of accumulating molecules and functions were preferentially lost in emerging archaeal lineages or preferentially accreted in the primordial emerging eukaryal-like lineages, and (iii) an organismal diversification epoch in which increasing numbers of lineage-specific variants of already existing molecules and functions appeared in an increasingly diversified tripartite world [20].

Our ToL showed that the most basal lineages belonged to crenarcheal hyperthermophiles of the orders Desulfurococcales and Thermoproteales. This observation supports the previously proposed thermophilic origin of the superkingdom [162] and extends it to diversified life. We note that the basal placement of Crenarchaota was also recently recovered in phylogenomic analyses of fold family domains [27], with roots that often included *Thermofilum pendens*. While clear grouping of recognized archaeal orders were evident in the tree, their relationships to each other were not so clear. The coherence of the Crenarchaeota and Euryarchaeota phyla originally identified using cultured strains on the basis of 16S rRNA [74] has been questioned by further addition of cultivars and environmental samples and by analysis of other molecules [111]. In contrast with Crenarchaeota, the Euryarchaeota has failed to represent a phylogenetically coherent group and has biological signatures related to Korarchaeota. However, new biological signatures of the archaeal groups and more widely encompassing phylogenetic analyses promise more clear definitions [163]. In our case, the ToL showed lack of coherence of both crenarchaeal and euryarchaeal microbes. However, it revealed groupings of archaeal orders, showed the

postulated close links between Sulfolobales and Thaumarchaeota, and included Korarchaeota within the euryarchaeal groups [163].

The non-thermophilic origin of Bacteria

Our ToLs failed to support a thermophilic origin for bacteria. This result is consistent with a number of recent studies (e.g. [164]) and challenges the canonical reference tree derived from 16S rRNA. While there is no general consensus for the branching order of bacterial phyla, trees generated from conserved 16S rRNA sequences have been rooted in Thermotogae and Aquificae, both of which include thermophilic and hyperthermophilic bacteria [165,166]. The most convincing support for the ancestral nature of thermophilic bacteria is the presence of the enzyme reverse gyrase that is found only in thermophiles (including thermophilic Archaea) [167]. This enzyme harbors two domains, an helicase-like domain and a Type I topoisomerase domain that is primarily responsible for positively supercoiling the DNA [168,169]. A recent evolutionary analysis of protein fold families however suggested a relatively late origin for the helicase-like domain [27] while the Type I topoisomerase domain was proposed to have been acquired from Archaea via HGT [167]. Many other studies based on different sets of genes and proteins also failed to recover the thermophilic rooting and sister relationship between Thermotogae and Aquificae [170-172]. Another study that focused on only the highly conserved and slow evolving sites of 16S rRNA revealed that both Thermotogae and Aquificae emerged later in evolution together with mesophiles (e.g. Fusobacteria), suggesting a secondary adaptation to life for the bacterial superkingdom [164]. The basal appearance in our ToLs of the anaerobic rod-shaped Bacteroidetes and some members of the PVC superphylum (Verrucomicrobia) is also compatible with the findings of Brochier and Philippe (2002) [164]. The phylum occupied deep positions in their tree, not far away from Planctomycetales, aquatic bacteria that often engage in parasitic relationships (and were excluded in our analysis). Remarkably, we found that the most basal orders of bacterial microbes in our ToL exhibited the highest level of reticulation that was observed ($\delta = 0.39$), which were derived from network reconstructions (Table 2.2). This suggests that HGT-like processes may have been important determinants in the emergence of the bacterial superkingdom. We conclude that the ancestor of Bacteria was more likely a mesophile that adapted to warm but comfortable environments that were becoming common on Earth about 2.1 billion years ago [38].

A close relationship between Plants and Metazoa

Within the strong monophyletic Eukarya, groups exhibited minimal trends of reticulation (δ = 0.16-0.28; Table 2.2) and main eukaryal kingdoms formed cohesive groups with taxa in the individual groups well positioned. Remarkably, the ToL of functionomes recovered again the close relationship of Metazoa and Plants that was obtained in previous phylogenomic analyses of domain structures (e.g. [20]) and domain interactomes [43]. The relationships of the fungal, plant and animal groups are the object of ongoing controversy as these have been consistently poorly resolved in sequence-based phylogenetic analyses [111]. This probably stems from a rather explosive radiation of eukaryotic crown taxa and phylogenetic reconstruction problems imposed by long-branch attraction and a 'Felsenstein's zone' defined by short internal branches followed by long edges in trees derived from sequences [173]. The congruent and well-supported relationship of plants and animals identified in the phylogenomic study of entire functionomic repertoires is therefore very encouraging and challenges the proposed fungal-animal split.

Advantages and limitations of GO terms as phylogenetic characters

In this study, we introduce a novel way of reconstructing organismal phylogenies built directly from the genomic ontological annotations. The choice of GO_{TMF} terms as phylogenetic characters carries several advantages over traditional phylogenies and few limitations that need to be addressed. The advantages include, but are not limited to: (i) GO_{TMF} terms portray organismal physiology and truly approximate the reconstruction of ToLs. (ii) GO_{TMF} terms represent a class of molecular characters that are more robust than amino acid or nucleotide site characters in sequence alignments. Sequence sites are prone to substitutions and suffer from high mutation rates [37]. In contrast, substitution of a molecular function into another function is rare. (iii) GO_{TMF} terms serve as informative tools to describe both the very deep and very derived organismal relationships. For example, the ancient GO_{TMF} terms that are evolutionarily conserved (e.g. ATP binding, structural constituent of ribosome) are highly abundant and widely distributed in living organisms [88]. This highlights the conserved nature of GO_{TMF} terms and their power to reliably describe deep relationships. In contrast, recently acquired GO_{TMF} terms by gene duplication or positive selection (e.g. diphosphokinase activity, coenzyme synthase activity) are less abundant and serve as useful tools to dissect the very derived branches of the ToL. Therefore, utilizing the genomic abundance of GO_{TMF} terms as phylogenetic characters

increases the resolution in both the very deep and derived branches of the ToL and enables reconstruction of reliable phylogenies. (iv) GO_{TMF} terms empower phylogenetic analysis by considering functional conservation. For example, the *Ly49* gene family in mice and *KIR* family in humans are sequentially non-homologous but both activate natural-killer cells of the immune system and trigger defensive mechanisms in a similar manner [174]. This represents a case of functional conservation that cannot be studied with molecular sequences. GO_{TMF} terms are advantageous in this regard as they account for the physiological responses of organisms and the genomic abundance value of molecular functions can be used to both group and differentiate organisms. (v) The impact of non-vertical evolutionary processes that can complicate traditional sequence-based phylogenies appears to be very minimal in our phylogenies.

With respect to limitations associated with the choice of GO_{TMF} terms as phylogenetic characters, we note that GO characters could well be interdependent. For example, a particular molecular function may be a consequence of another function and thus would require co-occurrence. However, this is a natural outcome of studying the evolution of entire systems (i.e. organisms), as individual parts in systems (GO terms in this case) are always dependent on other parts. This same problem exists for example when using gene, genome or concatenated gene sequences to build ToLs. While we have not yet explored or quantified the effect of co-occurrence of molecular functions, our paper lays foundation for functionomic network studies. Another possible limitation that is shared with sequence phylogenies is that the accuracy of the ToLs reconstructed in this study can suffer from individual GO_{TMF} terms harboring different evolutionary histories, especially because evolution of molecular functions depends on functional constraints. While incompatibility between phylogenetic characters decreases the accuracy of a tree topology, many previous studies have shown that multi-gene phylogenies are more robust than single-gene phylogenies. This indicates that the use of a large number of genes increases the amount of phylogenetic signal and overwhelms the problem of phylogenetic heterogeneity (summarized for genes in [175]). Consequently, the ToLs that were reconstructed by analyzing all available molecular functions should be considered robust against phylogenetic noise resulting from GO_{TMF} term interdependency and heterogeneity.

In this study, we used GO terms without reference to their evidence codes. As a result, our dataset included both manually and electronically curated GO terms. We have previously shown that tree topologies are robust against the difference of evidence codes and thus this

should not significantly affect our interpretations [88]. Finally, we expect functional annotations of genes to undergo revisions as more genomes are being sequenced. Thus it is possible that few GO_{TMF} terms sampled in this analysis are later classified as parent terms for some other terms. Therefore, our phylogenies and interpretations rest on current GO definitions and caution the reader to focus on general trends in our data rather than specific numbers, which are expected to change. However, we assume that global patterns described in our study will remain unaffected with an increase in genomic data.

Conclusions

In this study we introduce the reconstruction of trees of cellular life that describe the evolution of functionomes. These phylogenies are built directly from genomic ontological annotations that portray organismal physiology and truly approximate the construction of trees of organisms. Remarkably, our methodology recovered the tripartite nature of the living world heralded by the biological school of Carl Woese and the very ancient and thermophilic origin of Archaea embodied in multiple (paraphyletic) branching patterns of archaeal lineages appearing at the base of the ToL. The early rise of Archaea is not only compatible with several lines of molecular evidence we previously discussed but also supports paleobiological claims of early archaeal lipids and methanogenic activity linked to the fossil record [176-178] and the early archaeal role in biogeochemical processes [179]. The analyses also recovered a non-thermophilic origin for the bacterial superkingdom and a close relationship between Metazoa and Plants that excluded Fungi, dissecting a long-standing controversy associated with the trichotomy of crown eukaryotic taxa. Our results agree with a theoretical framework in which lineages evolve unique trade-off solutions among three strategies, economy, flexibility, and robustness [64]. This framework places evolving lineages in a 'persistence triangle' supported by protein domain structure and many other lines of evidence. Within the triangle, Archaea and Bacteria gravitate towards the triangle's economy vertex and arise very early in evolution, with Archaea biased towards robustness mainly due to very early adaptations to the thermophilic habitats of early Earth. Protista in turn occupy a saddle manifold that separates akaryotic microbes from multicellular organisms. According to this framework, the manifold was historically defined by the viscosity of water, which sets a critical barrier to organism size (100 mm) and possible trade-off solutions that unfold towards the economy vertex in microbes and delimit positive feedback loops towards flexibility and robustness in higher organisms. In our study, we also evaluated the effects of parasitic taxa (reductive evolution) and the functions of HTP characters (HGT) and suggest that they should be excluded for reliable interpretations. We conclude by proposing that functionomic data are useful and reliable additions to the toolkit of molecular features used for phylogeny reconstruction. The new ToLs that describe the evolution of functionomes reveal deep phylogenetic relationships with considerable explanatory power for the deep evolutionary study of cellular species. The new methodology can also yield novel insights into the evolution of

molecular functions in genomes, since phylogenetic characters describing potentially interesting molecular functions can be traced along the branches of the ToL.

Figures

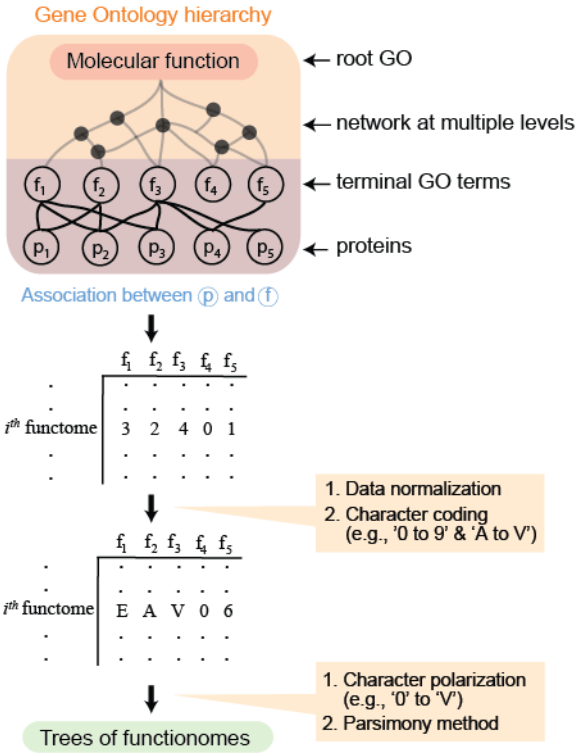

Figure 2.1 Overview of the phylogenomic methodology. The GO hierarchy with multiple levels associates proteins to biological, cellular and molecular roles. The genomic abundance counts of terminal GO terms (corresponding to molecular functions) were used as phylogenetic characters to describe the evolution of hundreds of functionomes (i.e. repertoire of functions). The abundance matrix was normalized and polarized to resolve compatibility issues with the phylogenetic reconstruction software PAUP*. Maximum parsimony was used to search for the best possible tree and to reconstruct trees of cellular life built directly from the ontological census of molecular functions.

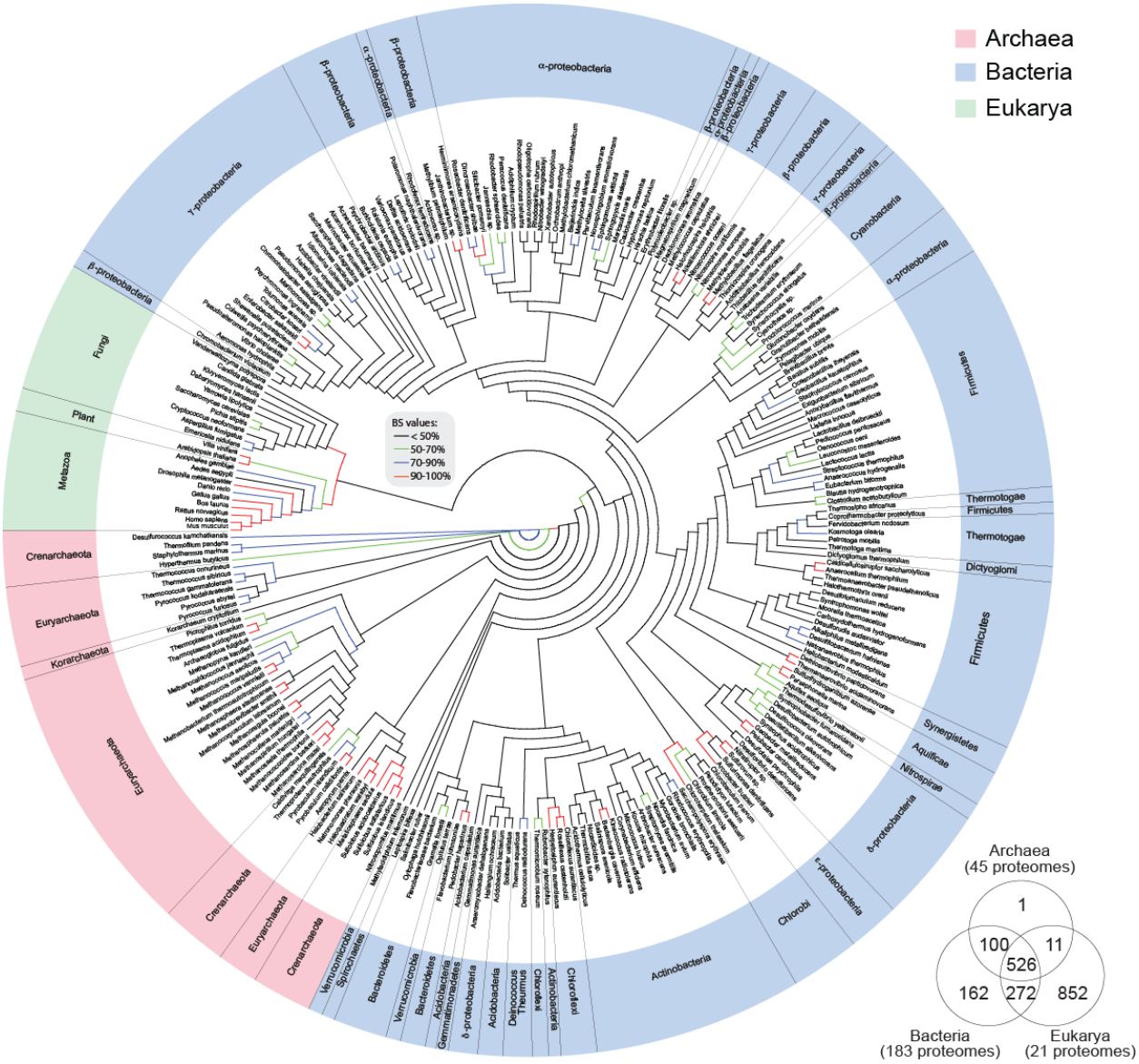

Figure 2.2 Tree of cellular life derived from the *non-HGT* functionome dataset. One optimal most parsimonious tree describing the evolution of 249 free-living functionomes built from the genomic census of 1,924 terminal GO terms that were not influenced by HGT (1,843 parsimony informative characters; 87,897 steps; CI = 0.1342; RI = 0.5798; g_1 = -0.839). Terminal leaves of Archaea, Bacteria and Eukarya were labeled in pink, blue and green, respectively, while major phyla and domains are also identified. Edges were colored according to BS values. The Venn diagram at the top describes the sharing patterns of GO_{TMF} terms between the three superkingdoms.

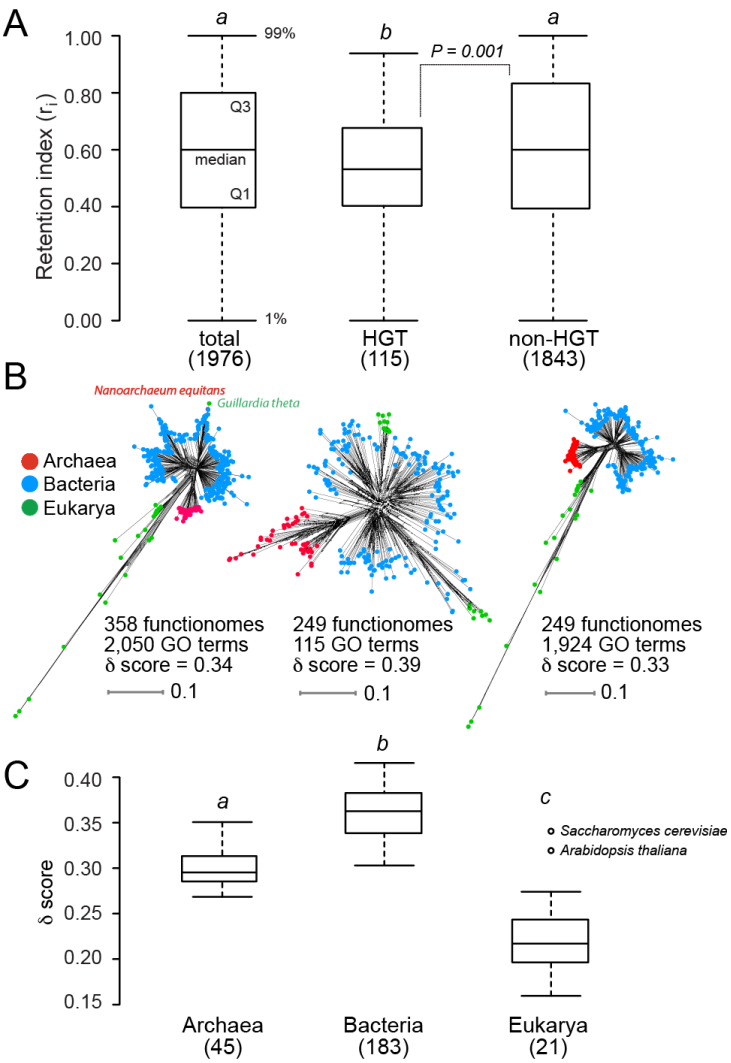

Figure 2.3 Reliability of phylogenomic trees and the evolutionary model. A) Boxplots comparing the fit of characters between trees reconstructed using various character sets. Mean, median and quartiles are identified. *P*-values are indicated for individual comparisons. Numbers in parentheses represent total number of parsimony informative characters for which r_i values were available. Boxplots headed by different letter are statistically significantly different**.** Statistical significance was evaluated using Student's unpaired two-tailed t-test at 95% confidence level. **B)** Phylogenomic networks generated for *total*, *HGT* and *non-HGT* datasets. Terminal nodes of Archaea, Bacteria and Eukarya were labeled in red, blue, and green, respectively. **C)** Boxplots comparing the distribution of δ-scores in the three superkingdoms. Outliers are labeled. Numbers in parenthesis indicate total number of free-living functionomes in each group. All the comparisons are significant at 0.05.

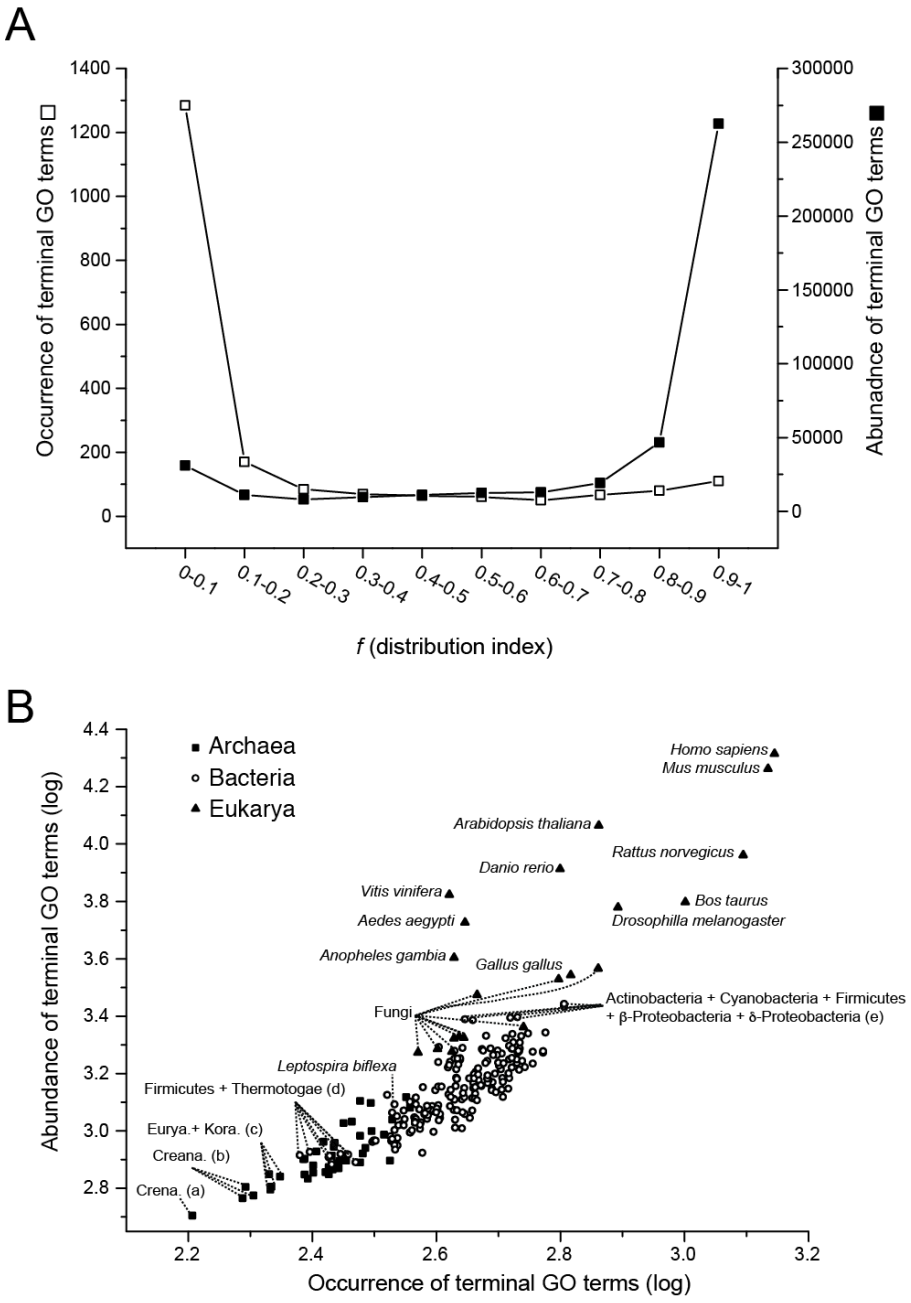

Figure 2.4 Relationship between abundance and occurrence. A) Occurrence and abundance values for terminal GO terms plotted against the distribution index (*f*; number of functionomes encoding a GO_{TMF} term / total number of functionomes). **B)** Abundance and occurrence counts plotted against each other for a number of functionomes. Both values are positively correlated. Axes are in logarithmic scale. (a) Crenarchaeaota (*Desulfurococcus kamchatkensis*); (b) Crenarchaeaota (*Hyperthermus butylicus, Thermofilum pendens, Staphylothermus marinus*); (c) Eutyarchaeota (*Thermococcus onnurineus, Thermoplasma acidophilum, Thermoplasma volcanium*) and Korarchaeaota (*Korarchaeum cryptofilum*); (d) Firmicutes (*Anaerococcus hydrogenalis, Eubacterium biforme, Pediococcus pentosaceus, Lactobacillus delbrueckii, Oenococcus oeni, Streptococcus thermophilus, Coprothermobacter proteolyticus, Leuconostoc mesenteroides, Macrococcus caseolyticus*) and Thermotogae (*Thermosipho africanus, Kosmotoga olearia, Fervidobacterium nodosum, Thermotoga maritima*); (e) Actinobacteria (Streptomyces avermitilis, Saccharopolyspora erythraea), Cyanobacteria (*Anabaena variabilis*), Firmicutes (*Brevibacillus brevis*), β-proteobacteria (*Ralstonia eutropha*) and δ-proteobacteria (*Haliangium ochraceum*).

Tables

Table 2.1 Measuring the degree of monophyly with the Genealogical Sorting Index (GSI). The GSI values and significance levels with 10,000 permutated replicates were examined for phyla having at least five proteomes.

Superkingdom	Phylum (no. proteomes)	*free-living*	*non-HGT*	*info*	*rRNA*
Archaea	Crenarchaeota (16)	0.66**	0.80**	0.63**	0.60**
	Euryarchaeota (28)	0.86**	0.70**	0.77**	0.92**
Bacteria	Actinobacteria (17)	0.83**	0.88**	0.78**	0.87**
	Bacteroidetes (6)	1.00**	0.83**	0.13*	0.28**
	Chlorobi (5)	1.00**	1.00**	1.00**	1.00**
	Cyanobacteria (6)	1.00**	1.00**	0.17**	1.00**
	Firmicutes (33)	0.82**	0.82**	0.49**	0.72**
	Proteobacteria-alpha (31)	0.62**	0.66**	0.69**	0.70**
	Proteobacteria-beta (18)	0.36**	0.48**	0.53**	0.87**
	Proteobacteria-gamma (27)	0.59**	0.74**	0.69**	0.41**
	Proteobacteria-delta (11)	0.48**	0.51**	0.13*	1.00**
	Thermotogae (5)	0.80**	0.80**	0.66*	1.00**
Eukarya	Fungi (10)	1.00**	1.00**	0.68**	0.21*
	Metazoa (9)	1.00**	1.00**	0.79**	0.56**

* $P < 0.05$; ** $P < 0.01$

Table 2.2 Comparison of average δ-scores in major taxonomic groups of superkingdoms.

Classification	Superkingdom	No. of taxa	δ-score
Euryarchaeota-Methanococci	Archaea	4	0.28
Euryarchaeota-Methanobacteria	Archaea	3	0.29
Euryarchaeota-Thermococci	Archaea	6	0.29
Crenarchaeota-Sulfolobales	Archaea	4	0.29
Crenarchaeota-Thermoproteales	Archaea	5	0.30
Euryarchaeota-Methanomicrobia	Archaea	9	0.30
Crenarchaeota-Desulfurococcales	Archaea	4	0.31
Euryarchaeota-Archaeoglobi	Archaea	1	0.31
Euryarchaeota-Thermoplasmata	Archaea	3	0.32
Euryarchaeota-Methanopyri	Archaea	1	0.32
Korarchaeota	Archaea	1	0.34
Euryarchaeota-Halobacteria	Archaea	3	0.34
Thaumarchaeota	Archaea	1	0.35
Thermotogae	Bacteria	5	0.31
Dictyoglomi	Bacteria	1	0.32
Synergistetes	Bacteria	2	0.33
Firmicutes	Bacteria	33	0.34
Nitrospirae	Bacteria	1	0.35
Proteobacteria-beta	Bacteria	18	0.35
Chlorobi	Bacteria	5	0.35
Aquificae	Bacteria	3	0.36
Proteobacteria-alpha	Bacteria	31	0.36
Proteobacteria-gamma	Bacteria	27	0.36
Deinococcus-Thermus	Bacteria	2	0.37
Proteobacteria-epsilon	Bacteria	4	0.37
Cyanobacteria	Bacteria	6	0.37
Proteobacteria-delta	Bacteria	11	0.38
Chloroflexi	Bacteria	4	0.38
Spirochaetes	Bacteria	1	0.38
Actinobacteria	Bacteria	17	0.38
Acidobacteria	Bacteria	3	0.39
Bacteroidetes	Bacteria	6	0.39
Verrucomicrobia	Bacteria	2	0.39
Gemmatimonadetes	Bacteria	1	0.39
Chordata-Mammals	Eukarya	3	0.16
Chordata-Primates	Eukarya	1	0.17
Fungi-Basidiomycota	Eukarya	1	0.20
Chordata-Birds	Eukarya	1	0.22
Chordata-Fish	Eukarya	1	0.23
Arthropoda	Eukarya	3	0.23
Fungi-Ascomycota	Eukarya	9	0.24
Plants-Streptophyta	Eukarya	2	0.28

CHAPTER 3: COMPARATIVE ANALYSIS OF PROTEOMES AND FUNCTIONOMES PROVIDES INSIGHTS INTO ORIGINS OF CELLULAR DIVERSIFICATION[3]

Introduction

Tracing the evolution of extant organisms to a common universal cellular ancestor of life is of fundamental biological importance. Modern organisms can be classified into three primary cellular superkingdoms, Archaea, Bacteria, and Eukarya [165]. Molecular, biochemical, and morphological lines of evidence support this trichotomous division. While the three-superkingdom system is well accepted, establishing which of the three is the most ancient remains problematic. Initial construction of unrooted phylogenies based on the joint evolution of genes linked by an ancient gene duplication event revealed that, for each set of paralogous genes, Archaea and Eukarya were sister groups and diverged from a last archaeal-eukaryal common ancestor [81,180]. This 'canonical' rooting that places Bacteria at the base of the 'Tree of Life' (ToL) is still widely accepted despite the fact that many other paralogous gene couples produced discordant topologies and despite known technical artifacts associated with these sequence-based evolutionarily deep phylogenies [181,182]. As a result, reconstructing a truly 'universal' ToL portraying the evolutionary relationships of all existing species remains one of the most controversial issues in evolutionary biology. This in part owes to the shortcomings of available phylogenetic characters and tree optimization methods that suffer from important technical and conceptual limitations [37,102] and have failed to generate a consensus. It is further complicated by the fact that genetic material can be readily exchanged between species, especially akaryotes (i.e. Archaea and Bacteria that lack a nucleus) via horizontal gene transfer (HGT) [10,25,183]. Non-vertical evolutionary processes coupled with uncertainties regarding evolutionary assumptions greatly complicate the problem of reconstructing the evolutionary past.

Recently, ToLs reconstructed using conserved structural information of protein domains [28,43], their annotated functions (Chapter 2), and universal RNA families [84,85,89,106,151,152] provided new ways to root phylogenies. These studies identified

[3]This chapter has been publishded as manuscript in *Archaea* (see [256]). The final publication is available at http://www.hindawi.com/journals/archaea/2013/648746/. Authors retain the rights to reprint.

thermophilic archaeal species to be the most closely related to the primordial cells. These findings not only challenge the bacterial rooting of the ToL but also highlight the importance of employing reliable phylogenetic methods and assumptions when reconstructing deep evolutionary history [102].

Here we advance the structural and functional approach by providing a simple solution to the problem of phylogenetic reconstruction. We argue that basic quantitative and comparative genomic analyses that do not invoke phylogenetic reconstruction are sufficient to resolve the tripartite division of cells and sketch their history. Our comparative approach involves the analysis of how superkingdoms, and their organismal constituents, relate to each other in terms of global sharing of genomic features. The genomic features we selected are entire repertoires of molecular structures and functions (collectively referred to as traits from hereinafter). They define two specific genomic datasets. The *structure* dataset encompasses the occurrence and abundance of 1,733 fold superfamily (FSF) domains in 981 completely sequenced proteomes. FSF domains were delimited using the Structural Classification of Proteins (SCOP ver. 1.75), which is a manually curated database of structural and evolutionary information of protein domains [33,34]. The FSF level of the SCOP hierarchy includes domains that have diverged from a common ancestor and are evolutionarily conserved [5,29]. In comparison, the *function* dataset describes the occurrence and abundance of 1,924 gene ontology (GO) terms [58,59] in 249 functionomes (Chapter 2).

We note that the global set of FSFs portrays the entire structural repertoire of organisms and that the repertoire of GO terms portrays their true physiology. Both provide useful information about species diversification. We restricted our analyses to include only structures and functions as they are more conserved than gene sequences and permit deep evolutionary comparisons [20,35,88]. In contrast, nucleotide sequences are susceptible to higher mutation rates and are continuously rearranged in genomes to yield novel domain combinations and molecular functions [37]. In other words, loss of an FSF domain structure or molecular function is much more costly for cells as it sometimes involves loss of hundreds of genes that have accumulated over long periods of evolutionary time to acquire a new structure or molecular activity. This is compounded especially for traits that are very ancient as they had more time to multiply in genomes and increase their genomic abundance [27,38]. Thus molecular structure

and function remain preserved in cells for relatively longer periods and make reliable candidates for inferring deep evolutionary relationships.

Here we show that an analysis of trait distribution between superkingdoms, distributions between genomic repertoires of superkingdoms, and abundance counts allow dissection of historical (ideographic) patterns using a comparative ahistorical (nomothetic) method (Figure 3.1). Inspired by a comparative analysis of RNA families [184], we measured the strength of evolutionary association between superkingdoms as a function of patterns of sharing of individual traits (Figure 3.1). We note that our approach is sufficiently informative to make reliable inferences regarding different evolutionary scenarios of diversification adopted by the three superkingdoms. This approach falsifies widely accepted theories regarding the origin of diversified life (e.g. [185,186]) and the fusion [69] and hydrogen scenarios [67] of eukaryotic origins, more than supporting any. This exercise then prompts validation by phylogenetic tree reconstruction, which we have reported previously (see [14,20,27,38]). In light of these considerations, the comparative exercise provides an easy-to use and reliable alternative to otherwise complicated phylogenetic tree reconstruction methods. These analyses carry the potential to yield significant insights into the evolution of cells and, if carefully interpreted, provide strong arguments in favor of the rooting of the ToL in Archaea and embedded canonical pattern of FSF and GO innovation.

Methods

Data retrieval and manipulation

FSF domain assignments for 981 completely sequenced proteomes were extracted from local MySQL installation of SUPERFAMILY ver. 1.75 database [40] using a stringent *E*-value cutoff of 10^{-4} [42]. The SUPERFAMILY database assigns structures to protein sequences using profile hidden Markov models (HMMs) searches that are superior in detecting remote homologies [41]. The dataset included 652 bacterial, 70 archaeal and 259 eukaryal proteomes encoding a total repertoire of 1,733 significant FSF domains. In this study, FSFs were identified using SCOP alphanumeric identifiers (e.g. c.37.1, where c represent the class of domain structure [α, β, α+β, α/β, etc.], 37 the fold, and 1 the FSF). This constituted the *structure* dataset.

To prepare the *function* dataset, we downloaded the Gene Ontology Association (GOA) files for 1,595 organisms from the European Bioinformatics Institute (http://www.ebi.ac.uk/GOA/proteomes). These files were filtered to exclude strain-level and parasitic organisms and then subjected to a 50% GO coverage threshold (i.e. number of gene products annotated to GO terms divided by the total number of gene products) to ensure high quality annotations (Chapter 2). In this study, we only sampled terminal-level GO terms from the GO molecular function hierarchy (simply referred to as GOs or functions from hereinafter), as they represent the highly-specialized functional annotations and approximate the molecular activities of cells (which are evolutionarily informative) [88]. We further excluded GOs that were likely candidates of HGT by scanning the total set of 2,039 terminal GOs in our dataset against proteins listed in the horizontal gene transfer database (HGT-DB) [134]. This allowed the exclusion of 115 potentially HGT-derived GOs. The final *function* dataset included 249 free-living functionomes from 183 Bacteria, 45 Archaea, and 21 Eukarya encoding a total set of 1,924 GOs.

Genomic census of traits

We conducted a genomic census for both *structure* and *function* datasets by calculating the occurrence (presence/absence) and abundance (redundant counts) of traits in all proteomes and functionomes. These data matrices were then scanned to generate Venn diagrams and boxplots displaying patterns of traits sharing both between and within proteomes and functionomes of superkingdom groups.

Calculating the spread of traits in proteomes and functionomes

The spread of each trait in a superkingdom was calculated by an *f*-value indicating the number of proteomes/functionomes harboring a trait divided by the total number of proteomes/functionomes in that organismal group. The *f*-value approaches one for ubiquitous traits but is lower for those that are less widely distributed.

Estimating the evolutionary age of traits

We used a relative time scale to pinpoint the origin of FSFs in molecular evolution. This scale was defined by node distance (*nd*) as calculated from a phylogenetic tree of FSF domains (see [20,29] for practical details). Technically, *nd* is the distance of a particular trait from its position on the phylogenetic tree to the root node. It is given on a scale from 0 (the most ancient or root node) to 1 (highly derived or terminal node). Biologically, it reflects the evolutionary age of an FSF relative to other FSFs. *nd* has been successfully used in the past to describe important events in the evolution of cells (e.g. [20,27]) and could be considered a reliable proxy to estimate the origin of molecular traits in organisms.

Results

Identifying vertical traces

Venn diagrams can be used to demonstrate the evolutionary sorting of FSF and GO traits in the seven possible and mutually exclusive Venn taxonomic groups, ABE (i.e. present in all three superkingdoms), AB (present only in akaryotes), BE (present only in Bacteria and Eukarya), AE (present only in Archaea and Eukarya), and the three superkingdom-specific groups, A, B, and E (Figure 3.2). Remarkably, the majority of the traits (45% of total structures and 27% of functions) were present in all three superkingdoms, supporting the hypothesis of common ancestry (Figure 3.2). Since a ToL by definition is a nested hierarchy of taxonomies, we propose that elevated sharing of traits by a taxonomic group points towards an ancient 'vertical trace' indicative of divergence from a common ancestor. In turn, low numbers in a taxonomic group are indicative of other evolutionary processes besides lineage splitting, including reductive evolution, HGT, convergent evolution, differential loss, and secondary evolutionary adaptations.

The two-superkingdom taxonomic groups were most informative as each embodied a possible vertical trace and an evolutionary hypothesis of superkingdom origin. The number of traits in the AB, AE and BE taxonomic groups are therefore indicative of the strength of evolutionary association between akaryotes, Archaea and Eukarya, and Bacteria and Eukarya, respectively. Remarkably, and against intuition, the size of the AB and AE taxonomic groups was ~9 folds smaller than that of BE in the *structure* dataset (38 and 38 vs. 324) (Figure 3.2A). This trend was also recovered in the *function* dataset where BE significantly outnumbered both AB and AE (272 vs. 100 and 11) (Figure 3.2B). These important biases suggest an intriguing ancestral evolutionary link between Bacteria and Eukarya, very much as the large number of ABE traits suggests an ancestral link between all organisms.

While simultaneous gains of traits in both bacterial and eukaryal proteomes would be possible, the high sharing of structures and functions by the BE taxonomic group makes it parsimoniously unlikely and points instead to an evolutionary scenario in which the two superkingdoms diverged from a common ancestor. This is particularly supported by the findings that convergent evolution of structures is rare [122] and seems unlikely to occur at such high levels. We note that bacterial organisms are more intimately associated with eukaryotes, establishing many coevolving bacterial parasitic/symbiotic interactions with eukaryotic hosts;

this is in marked contrast with organismal interactions involving Archaea [187]. These interactions could foster the exchange of protein and functional repertoires between the organisms. However, the *function* dataset included only free-living GO-annotated organisms with the exclusion of HGT-acquired GOs and consequently was free from adaptive effects of either parasitic or symbiotic lifestyles. The dataset still showed the high representation of the BE group relative to the AB and AE groups (Figure 3.2B). In short, the very large size difference of BE compared to the AB and AE groups is an evolutionarily significant outcome that cannot be explained merely by parasitic/symbiotic processes.

Finally, the Venn diagrams show that Eukarya-specific traits always outnumbered Bacteria-specific and Archaea-specific counterparts, suggesting either an expansive mode of evolutionary growth of eukaryotic repertoires or a reductive mode in akaryotic counterparts, or both (Figure 3.2B). This is an expected result as eukaryotes encode a highly diverse and complex genome and are capable of carrying out many advanced molecular activities, especially those related to development and immunological responses. Based on our initial comparative genomic exercise, we put forth three preliminary conclusions, (i) all extant cells are related by common descent, (ii) Bacteria and Eukarya diverged from a mutual ancestor, and (iii) eukaryotes are significantly more complex than akaryotes in terms of numbers of unique traits.

Identifying horizontal traces

Venn diagrams simply describe global patterns of sharing in superkingdoms and cannot dissect how popular are traits in organisms of each superkingdom. In other words, the presence of a trait in a superkingdom does not necessarily imply that it was vertically inherited; this trait might only be present in few of its members. In such cases, acquisition of traits by non-vertical (e.g. HGT fluxes, convergent evolution) or confounding (e.g. differential loss that mimics HGT) evolutionary processes becomes more likely. To fully explore the extent to which these real or virtual 'horizontal traces' contribute to the development of the proteomes of organisms in superkingdoms and to further test the preliminary conclusions drawn from the Venn diagrams of Figure 3.2, we calculated the spread or popularity of FSF and GO traits in the organisms of superkingdoms, which we term *f*-value.

The *f*-value is simply the number of organisms in a Venn taxonomic group harboring a trait divided by the total number of organisms in that taxonomic group and in that superkingdom.

It is given on a relative scale from 0 (absent) to 1 (omnipresent). Using this simplistic approach, we first identified 17 FSFs (Table 3.1) and 26 GOs (Table 3.2) that were present in all proteomes and functionomes, respectively. This cohort of traits truly represents the 'universal' core that was present in the common ancestor of life, the urancestor, and was strongly retained by all of its descendants. These traits perform crucial and central metabolic and informational roles in cells such as ATP hydrolysis and ion binding, make up structural components of ribosomal proteins, and are involved in DNA replication and protein translational processes (Tables 3.1 and 3.2). Moreover, a total of 245 FSFs and 95 GOs had an $f > 0.90$ implying near-universal presence and suggesting reductive losses in the remaining 10% of the proteomes and functionomes (data not shown). This global analysis based on the popularity of traits in proteomes and functionomes suggests that the urancestor was especially enriched (structurally and functionally) in metabolic functions [20,38], and illustrate the power of *f*-value in dissecting traces of vertical vs. horizontal inheritance. Therefore we extended this analysis to the proteomes and functionomes of members of each of the seven taxonomic groups.

We first compared the spread of FSFs in the *structure* dataset using boxplot representations of *f*-value distributions (Figure 3.3A). Our assumptions are straightforward: high *f*-values and balanced *f*-distributions reflect vertical traces while low *f*-values and biased *f*-distributions echo horizontal (flux-loss) traces, respectively. The 786 ABE structures were distributed with the highest *f*-values and the medians increased in the order, Archaea (median $f = 0.6$), Bacteria (0.74), and Eukarya (0.90) (Figure 3.3A, ABE taxonomic group). The large number of ABE structures that was widespread in all three superkingdoms strengthens the hypothesis of life's common ancestry. The relatively lower median *f*-values in akaryotes (0.6 for Archaea and 0.74 for Bacteria vs. 0.90 in Eukarya) can be explained by genome reduction events that are known to occur with relatively high frequency in akaryotic microbes [20,86], and also manifest in the numbers of superkingdom-specific traits (Figure 3.2). The 38 AB structures were poorly but similarly distributed (median *f*-values = 0.14) in archaeal and bacterial proteomes, with archaeal structures exhibiting a tendency to become more widespread (longer tail) (Figure 3.3A, AB taxonomic group). This pattern supports the existence of a horizontal trace between akaryotes, with a weak bias in flux-loss between superkingdoms (note however that no common outliers could be detected). In contrast, the 38 AE structures spread were highly represented (median *f*-values > 0.94) in the organisms of corresponding superkingdoms (Figure 3.3A, AE

taxonomic group). Again, archaeal structures appeared more widely shared but also showed a longer tail indicative of possible flux-loss episodes. At first glance, this chimes for a strong vertical trace of the AE group that could rival that of the BE group. However, this may not be the case. The 324 BE structures were on average poorly represented in bacterial and eukaryal proteomes (median *f*-values < 0.15) (Figure 3.3A, BE taxonomic group). Their overall spread was relatively uniform, with a weak bias towards higher representation in Eukarya. However, 53 and 59 structures were widely shared by the proteomes of Bacteria and Eukarya ($f > 0.8$), respectively (shaded region in Figure 3.3A, BE boxplots). This subset of BE structures was numerically double that of the total set of the highly represented AE structures. Thus, the stronger vertical trace for BE structures continues to support a sister-group relationship between Bacteria and Eukarya and the early diversification of Archaea. We note that this inference is strengthened by the fact that we had 652 bacterial and 259 eukaryal proteomes in comparison to only 70 archaeal proteomes. Existence of any structure in such large number of genomes implies strong selective pressure and conservation of that trait. Finally, the sharing of superkingdom-specific structures was low in each superkingdom (median *f*-values = 0.01-0.34), with minimum average *f*-values for Bacteria and maximum for Eukarya (Figure 3.3A, A, B, and E taxonomic groups).

Remarkably, out of the 164 Bacteria-specific structures, none, but one, were present in >50% of the proteomes (Figure 3.3A, B taxonomic group). The absence of an expected homogenous distribution strongly suggests that the role of HGT and other homogenizing processes may be quite limited in shaping the evolution of bacterial proteomes. Eukaryal-specific structures were distributed with higher *f*-values (Figure 3.3A, E taxonomic group). The relatively low spread of superkingdom-specific structures suggests that these structures were acquired independently and after divergence from the last common ancestors of each superkingdom.

Inferences drawn from boxplots of the *function* dataset (Figure 3.3B) again supported the general conclusions derived from the *structure* dataset. The ABE distributions had high *f*-values, with those of Archaea (median $f = 0.24$) being considerably lower than those of Bacteria (0.57) and Eukarya (0.57) (Figure 3.3B, ABE taxonomic group). Bacterial and eukaryal distributions were remarkably homogenous, providing additional support to their recent divergence from a mutual ancestor. The median *f*-value in Archaea was lowest and could be explained by either high genome reduction events [20] or biases in the number of GO annotations for archaeal

genomes. GOs are more reliably and extensively curated for Bacteria and Eukarya, and this factor could reduce the number of overall detections in archaeal genomes. However, comparing distributions of the *function* and *structure* datasets show supporting results were consistent and suggest a limited impact of this possible shortcoming. Here, ABE distributions followed the pattern observed for FSFs and were therefore considered reliable. None of the AB, AE, and BE taxonomic groups showed balanced distributions (Figure 3.3B, AB, AE, and BE taxonomic groups). The AB taxonomic group harbored 100 GOs (~3 fold greater than corresponding structures) that were distributed with low popularity (Figure 3.3B, AB taxonomic group). In general, these functions were more abundant in Bacteria compared to Archaea and thus suggested that some molecular activities were laterally transferred from Bacteria to Archaea (confirmed below). The AE taxonomic group failed to strongly support AE distributions in the *structure* dataset. This group included only 11 GOs that were relatively more abundant in eukaryal proteomes (Figure 3.3B, AE taxonomic group). Finally, the BE taxonomic group also supported the increased prevalence of BE functions in eukaryal genomes compared to bacterial genomes (0.39 median vs. 0.03), indicating either horizontal trace effects or biases introduced by GO annotation schemes (Figure 3.3B, BE taxonomic group). However, the numbers of traits of the BE group were considerably greater than those of either the AB or AE groups and included a significantly large number of functions that were relatively widespread ($f > 0.8$) (Figure 3.3B, BE taxonomic group). This was in sharp contrast with patterns in either AB or AE taxonomic groups. The subset of highly represented BE functions is therefore the most likely trace of an ancient vertical signature that unifies Bacteria and Eukarya as sister-groups in the ToL. This trace is remarkably consistent with the patterns obtained in the *structure* dataset (Figures 3.2A, and 3.3A).

Finally, the superkingdom-specific functions were again distributed with low *f*-values. Archaea had only one unique GO that was present in 40% of the archaeal genomes (Figure 3.3B, A taxonomic group). In sharp contrast, there were 162 bacterial and 852 eukaryal-specific GOs. Bacterial functions again showed evidence of very limited spread in organisms (Figure 3.3B, B taxonomic group) challenging claims of widespread bacterial HGT. In turn, eukaryal functions were moderately widespread (Figure 3.3B, E taxonomic group). These results are in line with earlier inferences regarding late and independent acquisition of superkingdom-specific traits.

Identifying patterns of horizontal flux

Boxplot distributions provided useful clues regarding the divergence patterns of superkingdoms. However, they did not allow us to quantify the extent of horizontal vs. vertical inheritance. Therefore, we calculated a difference in the *f*-value for all traits in the AB, AE, and BE taxonomic groups. If the difference between *f*-values was > 0.6, the presence of the trait in both superkingdoms was considered the result of a probable HGT event. This threshold was set arbitrarily to include only those traits that were considerably more abundant in one superkingdom but scarcely present in the other. For example, the 't-snare proteins' superfamily (SCOP Id: a.47.2), which is abundantly found in yeast and mammalian cells and forms bridges to mediate intracellular trafficking [188], had an *f*-value of 0.996 in eukaryotes implying that it was ubiquitous. However, it was only present in one of the 652 bacterial proteomes examined (f = 0.001). This most likely is an example of structure gain via HGT that occurred in the direction from Eukarya to Bacteria.

Using this criterion, only one structure ('tRNA-intron endonuclease N-terminal domain-like' [d.75.1]) was acquired horizontally in Eukarya from Archaea in the AE taxonomic group, while 6 were transferred from Eukarya to Archaea. Similarly, only one FSF was laterally transferred to Bacteria from Archaea ('Sulfolobus fructose -1,6-bisphohsphatase-like' [d.280.1]) while none were acquired in reciprocity. Finally, Bacteria likely transferred 35 structures to eukaryotes while gained 52 in return. The rest 237 structures did not show significant deviations in terms of spread in these taxonomic groups and were possibly acquired vertically or gained independently in evolution.

In *function*, none of the GO traits were likely transferred to Bacteria from Archaea. However, 9 GOs were transfer candidates from Bacteria to Archaea. Perhaps the most interesting among these was the lateral acquisition of 'penicillin binding molecular activity' [GO:0008658] that was universally present in Bacteria but also present in 11% of the archaeal proteomes. Similarly, no molecular function was transferred to Eukarya from Archaea, while only one GO ('dolichyl-diphohphooligosaccharide-protein glycotransferase activity' [GO:0004579]) was gained. Finally, 4 molecular functions were likely transferred from Bacteria to Eukarya and 28 were gained in return. Overall, the inferred impact of horizontal transfer processes appeared quite limited and did not seriously invalidate our inferences. Moreover, horizontal contributions from Archaea to either Bacteria or Eukarya were minimal, which is consistent with the minimal sharing of traits described above (Figures 3.2, and 3.3). In comparison, both Bacteria and

Eukarya exhibited higher levels of vertical and horizontal inheritance of traits and indicated a much stronger evolutionary association, a conclusion intimated by likely ancient endosymbiotic events.

Identifying ancestral traits using abundance counts

Traits that are of ancient origin are expected to be present in greater abundance than those acquired recently. This is true because traits appearing earlier have more time to accumulate in genomes and to increase their representation [27,37]. Thus, high abundance of traits in a particular Venn taxonomic group is indicative of presence of relatively more ancient traits and an ancient origin. Therefore, genomic abundance can be used as one proxy to estimate the age of taxonomic groups.

We calculated the abundance of traits present in each proteome and functionome and represented these values in boxplot distributions (Figure 3.4). The median abundance value was highest for the ABE taxonomic group in both the *structure* (Figure 3.4A) and *function* (Figure 3.4B) datasets, again supporting that this group retains most of the urancestral traits that have relished maximum time to multiply and become abundant in modern proteomes and functionomes. The BE group always harbored traits in much greater abundance compared to the AB and AE groups (Figure 3.4). Finally, Eukarya-specific traits were significantly enriched in the eukaryal proteomes and functionomes and were detected in much greater abundance compared to the genomic abundance of either Archaea-specific or Bacteria-specific traits (Figure 3.4). This result confirms the existence of a strong vertical trace in modern cells in the direction from ABE to BE and to E. It is likely that eukaryotes retained the majority of the most ancient traits that were progressively lost in akaryal organisms, beginning in Archaea and manifesting much later in Bacteria. Previous phylogenomic analyses have confirmed strong reductive trends in the akaryal proteomes [14,20,27,86]. Evolution of Archaea has also been linked to genome reduction events that started very early in evolution and before the appearance of the BE taxonomic group [14,27]. However, the relatively late loss of traits in Bacteria is intriguing. Several bacterial species are known to have adapted a parasitic lifestyle following genome reduction [22]. Thus gene loss in Bacteria is likely an ongoing evolutionary process hinting towards a major secondary evolutionary transition. This was also manifested in the very poor spread of Bacteria-specific traits (Figure 3.3).

We provide evidence for late loss in Bacteria by closely examining the AE traits. The majority of the 38 AE FSFs and 11 GOs are enriched in informational functions (e.g. translation initiation, ribosomal proteins, DNA binding proteins, proteins involved in DNA replication; Tables 3.3 and 3.4). This result is consistent with existing knowledge. Indeed, Archaea and Eukarya are more related to each other in terms of informational processes, while Bacteria and Eukarya resemble each other metabolically [189]. Thus, the high popularity of AE FSFs could be due to biases attributed to late differential loss of structures in these functional categories. For example, the 11 AE GOs include crucial molecular functions such as 'DNA polymerase processivity factor activity [GO:0030337]' and 'tRNA-intron endonuclease activity [GO:0000213]'. The former is a regulator of the replication fork [190,191] while the latter is involved in processing tRNA introns [192]. Both of these activities could be linked to late losses in Bacteria, as they seem centrally important functions in cells. Therefore, while HGT, convergent evolution and co-evolution of BE traits seems less likely, we cannot rule out the possibility of extensive genome reduction in akaryal species.

Tracking the vertical trace

To further dissect the evolution of Venn taxonomic groups, we mapped the 1,924 terminal GOs to 16 level 1 parent GO terms. Figure 3.5 shows the distribution of terminal GOs, indexed by taxonomic group, in each of the 16 parent categories. This exercise confirmed the inferences drawn from earlier experiments and highlighted the direction of the vertical trace.

Remarkably, only ABE, BE, and E were enriched in level 1 molecular functions while the majority of the terminal GO terms could be identified as either 'catalytic activity [GO:0003824]' or 'binding [GO:0005488]' (Figure 3.5). This is an interesting result. A previous analysis by Kim and Caetano-Anollés [88] confirmed that these two molecular activities appeared first in evolution and were shared by all organisms. In comparison, the more derived molecular activities first appeared in the BE taxonomic group (e.g. 'structural molecule activity [GO:0005198]', 'nucleic acid binding transcription factor activity [GO:0001071]', and 'channel regulator activity [GO:0016247]), while the recent innovations occurred uniquely in Eukarya (e.g. 'receptor regulator activity [GO:0030545], 'translation regulator activity [GO:0045182]', 'metallochaperone activity [GO:0016530]', 'morphogen activity [GO:0016015]', and 'protein tag [GO:0031386]'). In contrast, none of the AB, AE, A, and B taxonomic groups uniquely

harbored a level 1 molecular function (Figure 3.5). Remarkably, a significant proportion of BE terminal GOs was devoted to the most ancient catalytic and binding activities (Figure C1). In comparison, 'transporter activity [GO:0005215]' was found to be over-represented in the AB group while AE was numerically much smaller (Figure C1). These findings strongly suggest the existence of a vertical trace from ABE to BE and finally to E (also supported by the *structure* dataset). Akaryal ancestors likely diverged from this trace by following paths towards genome reductions while eukaryotes enriched their repertoires by engaging in gene duplication events and exploring novel domain combinations [6,43].

Validating inferences with evolutionary timelines

To validate our ahistorical comparative approach, we unfolded the appearance of FSF and GO traits in evolutionary time (*nd*), while plotting their genomic abundance in each superkingdom. The historical analysis of FSF evolution (Figure 3.6) and GO terminal terms (data not shown) were congruent and revealed two clear patterns: (1) a pattern of ancient genomic loss embodying the early rise of the BE taxonomic group (red circles), which generally involved traits with abundance levels that were at least an order of magnitude higher than the levels of other taxonomic groups (e.g. AE and AB); and (2) a canonical pattern of appearance of superkingdom-specific traits that revealed the rise of early bacterial novelties followed by the joint appearance of unique novelties in Archaea and Eukarya. This historical analysis therefore supports the ancient vertical trace identified by comparative analysis that flows from the ABE group to the BE and E groups. These three groups were distributed with maximum abundance values in timelines indicating retention of large number of traits from the common ancestor. This vertical trace defines an ancient stem line of descent responsible for the early origination of archaeal lineages and bacterial novelties, which reconciles the canonical and archaeal rooting of the ToL.

The historical analysis however was unable to predict the canonical pattern, since the comparative analysis of trait distribution in Venn taxonomic groups, superkingdoms and organisms cannot accommodate competing hypotheses of rooting that manifest at different times in evolution. The plots of Figure 3.6 also revealed a marked increase in the abundance of FSFs late in eukaryal evolution, which can be explained by the remarkable development of multidomain protein structures and their associated functions [6,43]. The combinatorics of

domains and functions is the likely culprit of the biphasic patterns we observed when we focus on Eukarya.

Discussion

Our approach is simple (Figure 3.1). It does not involve computation of a sequence alignment or use of complex data matrices for phylogenetic reconstruction. Instead, it focuses on the census of molecular (structural and functional) traits in the genomes of modern cells. The fundamental principle of analysis is the use of trait distributions in Venn taxonomic groups to explain vertical evolutionary traces, the use of *f*-values to explain horizontal traces, and the use of trait abundance as a proxy for age. The sequential combination of these approaches dissects the most likely scenario of diversification of superkingdoms, without invoking a phylogenetic framework of analysis.

Our comparative genomic exercise shows evidence in favor of a common ancestry for cells and establishes the deep branching patterns of the ToL. The genetic complexity of Bacteria and Eukarya hints towards a strong and ancient evolutionary association between the two superkingdoms. This association is stronger than the associations of other superkingdoms. Our findings are also compatible with an evolutionary scenario in which Archaea emerged as the first superkingdom of life by diverging from a primordial stem line of descent that originated in the urancestor [20,27]. This line likely encountered extreme temperatures that affected its proteomic growth, hampering the acquisition of new molecular traits in those environments. Under such hostile conditions, the persistence strategy of the emergent archaeal cells was most likely survival rather than enrichment [64]. This explains why we observed the lowest number of traits in extant archaeal species. In contrast, both Bacteria and Eukarya shared a protracted co-evolutionary history. Their diversification occurred well after the primordial split of Archaea from the urancestral line. Bacteria followed a path towards exploring a diverse range of habitats, which enabled high rates of gene discovery. This explains the high numbers of unique bacterial traits that are unequally distributed among bacterial species. Bacterial species also engaged in genome reductive processes and simplified their trait representations. This probably occurred well after their divergence from the primordial stem line. Finally, eukaryotes evolved by (i) increasing the abundance of ancient traits (via gene duplications and domain rearrangements), (ii) discovering novel traits, or (iii) both. These findings falsify an evolutionary scenario of first appearance of bacterial cells [185] or the fusion and hydrogen hypotheses linked to the origin of eukaryotes [67,69], as none seem compatible with our data. However, we did not consider the

roles that viruses may have played during cellular evolution. Viruses are known to contribute to the genetic diversity of cells and are believed to be very ancient [14,193-195]. We aim to accomplish this task in the near future.

Genome reduction is an ongoing evolutionary process that often triggers lifestyle transitions in cells (e.g. from free-living to intracellular parasites [22]). We propose that genome streamlining played a key role in the evolution of akaryotes, especially Archaea. Our data show that the BE taxonomic group was enriched in molecular traits compared to the relatively poor representations of FSFs and GOs in the AB and AE groups (Figure 3.2). In fact, phylogenomic analysis revealed that the BE group appeared very early in evolution and was correlated with high abundance levels of BE FSFs in bacterial and eukaryal proteomes (Figure 3.6). These findings were taken as an indication of loss of traits in Archaea that occurred very early in evolution. While it can be argued that such losses could have occurred much later in archaeal lineages and after their diversification from Bacteria, our comparative and phylogenetic data indicate that this may not be very likely. The loss of ancient traits late in evolution is phylogenetically costly as it implies loss of many genes and proteins that have accumulated during the course of evolution to perform a particular molecular task. In comparison, loss of ancient traits early in evolution is more parsimonious and complies with the principle of continuity. An alternative explanation, however, could be confounding effects of HGT processes. However, it was shown recently that a large number of ribosomal proteins were unevenly distributed in archaeal species [22,196]. Because ribosomal proteins are assumed to be refractory to HGT, their patchy and uneven distribution in archaeal lineages is better explained by differential loss from a more complex archaeal ancestor. Taken together, these findings strongly suggest that primordial reductive evolutionary processes have tailored archaeal evolution.

When placed along evolutionary timelines of trait innovation (Figure 3.6), Venn taxonomic groups uncovered a remarkable pattern that could not be dissected with the comparative genomic approach. This hidden pattern embodies the primordial rise of Bacteria-specific traits followed much later by the concurrent appearance of Archaea-specific and Eukarya-specific innovations. This important succession supports the ‘canonical’ rotting of the ToL in which Bacteria occupy the most basal positions while Archaea and Eukarya emerge as derived sister-groups [81,180]. From a cladistics perspective, traits unique to a superkingdom are autapomorphies, derived features that are unique to terminal groups. These autapomorphies

cannot be used to reconstruct trees in phylogenetic analysis or dissect the alternative evolutionary scenarios of our comparative genomic approach. In comparison, FSFs and GOs that are shared by any two superkingdoms reflect synapomorphies (shared and derived features) that allow both historical (phylogenetic) and ahistorical (comparative) inferences. We note that traits uniquely shared by any two superkingdoms can arise either by the gain of the feature in two superkingdoms or by the loss in one. Abundance levels and *f*-distribution patterns support the latter scenario, especially if the loss involves an ancient trait. Thus, an early primordial loss of FSFs and GO synapomorphies in Archaea embeds later on the early gain of autapomorphies in Bacteria.

The hidden canonical pattern of Figure 3.6 was already reported in an exhaustive structural phylogenomic exploration of domain evolution at fold and FSF levels of structural abstraction [20], which prompted the definition of three epochs in the evolution of proteins and the organismal world and a number of hypotheses of origin. In the first *'architectural diversification'* epoch, the emerging organismal community accumulated a rich toolkit of protein structures and functions. This communal world resembled the ancient world of multi-phenotypical pre-cells proposed by Otto Kandler [197] that inspired Carl Woese's more advanced scenarios of early cellular evolution [198]. However, and in contrast with the simple cellular systems sought by Kandler and Woese, the pre-cell molecular make up that was inferred from our phylogenomic analysis was extremely rich in complex structures and functions [38]. This richness expresses today in the sizable number of structures and functions that are shared by all superkingdoms and are revealed by our comparative exploration. Towards the end of the architectural diversification epoch, the pervasive loss of domain structures in subgroups of the urancestral pre-cell population resulted in primordial archaeal grades, groups of diversifying organisms in active transition that were at first unified by the physiological complexity of the urancestral community but later on gained the cellular cohesiveness needed to establish lineages and true patterns of organismal diversification. While it may prove difficult to establish the time when these 'thresholds' (*sensu* [198]) were crossed by the primordial archaeal grades as these were stemming from the urancestral stem line, the early process of reductive evolution left deep historical signatures in the make up of the archaeal organisms that are embedded in the timelines of domain structures [20].

The second *'superkingdom specification' epoch* brought the first Bacteria-specific domain structures and later on the concurrent appearance of Archaea-specific and Eukarya-specific structures. This canonical pattern of appearance of superkingdom-specific structures, which unfolded in the absence of early and major reductive evolutionary tendencies, signals a time in which the emerging superkingdoms were being molded by innovation. During this epoch, grades turned into clades and the pre-cell 'swap shop' strategy was gradually replaced by organismal cohesiveness. Marked decreases in *f*-values during this time suggested that lineage sorting occurred more frequently in the growing number of lineages.

Finally, in the *'organismal diversification' epoch,* commitment to strategies and lifestyles enhanced even further the divide between superkingdoms and weakened the contribution of the stem line of descent. Two forces of particular significance play crucial roles during this final epoch, the combinatorial use of domains as modules in multidomain proteins of Eukarya [6,43] that is responsible for the high abundance levels and the biphasic patterns of Figure 3.6 and the HGT-driven combinatorial exchange of protein repertoires in lineages of Bacteria [20] that minimizes trait distribution in Figure 3.3.

We conclude by emphasizing that our comparative genomic inferences have been ratified previously by phylogenetic tree reconstructions [5,20,27,28,43,84,89] and thus establish the power of our methodology. However, our analysis depends upon the accuracy and sampling of structures and functions and the reliability of the datasets. The *function* dataset, in particular, is dependent upon the stability of GO annotations and is biased towards eukaryal organisms that are more carefully annotated. To minimize this factor, we sampled 183 bacterial and 45 archaeal functionomes in comparison to only 21 eukaryotes. Despite the huge number of akaryal functionomes in our dataset, we were still able to highlight the incredible enrichment of eukaryal repertoires. Moreover, inferences drawn from *function* were in agreement with *structure* and both should be considered reliable. While tracing back evolutionary history from the present to the first cell is a complex problem, inferring the patterns of species diversification by comparing the use and reuse of molecular traits in extant cells must be considered a robust inferential approach that is free from many of the external assumptions and technical problems faced when reconstructing phylogenetic trees. The only shortcoming may be one of interpretation, which we here showcase with the scenarios of origin we have discussed. However, we have tried to restrict our statements to scenarios that seem most compatible with given data. An example is using a

threshold of 60% difference in the popularity of traits to detect HGT-derived structures and functions. This criterion was set arbitrarily to identify only the most likely HGT-transfers but may have resulted in failure to detect some of the true HGT-acquired traits, especially for those where both inter-superkingdom and intra-superkingdom transfers occurred rapidly. Although such events are less likely, they may still be occurring. However, detection of such transfers is a hard problem and cannot be reliably confirmed without experimental evidence. Given the conservation levels of structural and functional traits and the relatively poor repertoire of likely HGT-acquired features, we safely assume that this factor did not seriously compromise our inferences. Finally, our approach is a systematic application of morphological analyses that were initially used to classify higher-order organisms. Future work should be focused on advanced applications of our approach in hope to come to a consensus regarding the evolution of cells.

Conclusions

We inferred evolutionary patterns by examining the spread of molecular features in contemporary organisms. The analysis revealed a common origin for all cells, the early divergence of Archaea and a sister relationship between Bacteria and Eukarya. Archaeal evolution was primarily influenced by genome reduction while that of Bacteria by two contrasting phases, (i) a period of early innovation that coincides with the rise and diversification of the bacterial superkingdom, and (ii) a post-divergence period of this lineage exhibiting relatively late genome reduction events. The branch leading to modern eukaryotes was minimally affected by reductive pressure and retained the majority of the ancestral traits. Eukaryotes further enriched the genomic abundance of these traits by engaging in gene duplication and domain rearrangement processes and by discovering novel structures and molecular activities. Traces of all of these events could be reliably detected in modern proteomes and functionomes. In particular, a strong vertical trace from the urancestor to the stem line unifying Bacteria and Eukarya and the ancestor of Eukarya could be inferred. This strong vertical trace strongly supports the existence of a stem line of descent, from which all three superkingdoms emerged, very much in line with Kandler's ideas of an aboriginal pre-cellular line of early biochemical evolution that was undergoing cellularization [197]. Finally, non-vertical evolutionary processes seemed to have played only limited roles during defining steps of cellular evolution. The comparative framework enables exploration of deep evolutionary histories without invoking tree reconstruction algorithms and external hypotheses of evolution. This approach is in line with various published phylogenetic analyses and provides strong support to theories favoring an archaeal origin of diversified life.

Figures

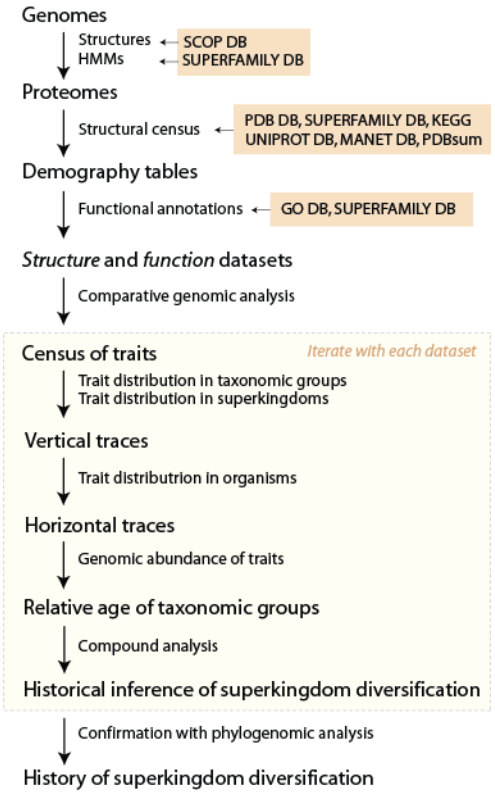

Figure 3.1 Overview of the comparative proteomic and functionomic methodology. Proteomes and functionomes were scanned for the occurrence and abundance of FSFs and GO terms (i.e. traits). This information was represented in data matrices that were analyzed for trends of trait sharing and traces of vertical and horizontal inheritance. Inferences were drawn regarding superkingdom diversification and were confirmed with previously published phylogenetic studies.

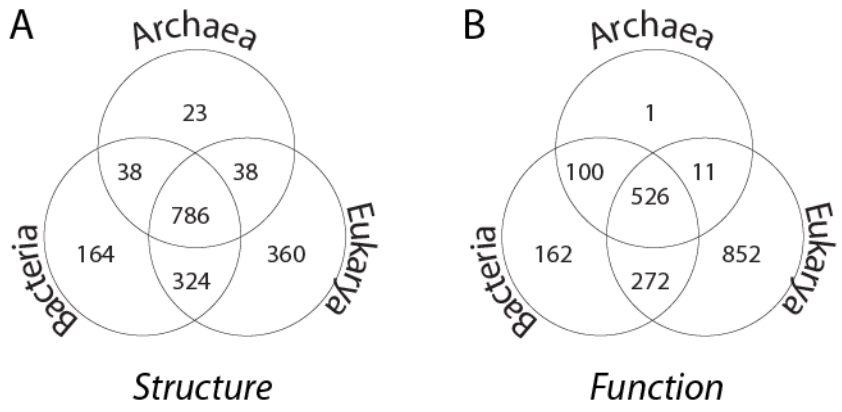

Figure 3.2 Global trends of trait sharing in Venn taxonomic groups. A) Venn diagram displaying the distribution of 1,733 FSF domains in 981 completely sequenced proteomes sampled from 652 Bacteria, 70 Archaea, and 259 Eukarya. This constituted the *structure* dataset. **B)** Venn diagram displaying the distribution of 1,924 terminal-level GOs in 249 free-living functionomes corresponding to 183 Bacteria, 45 Archaea, and 21 Eukarya. This constituted the *function* dataset.

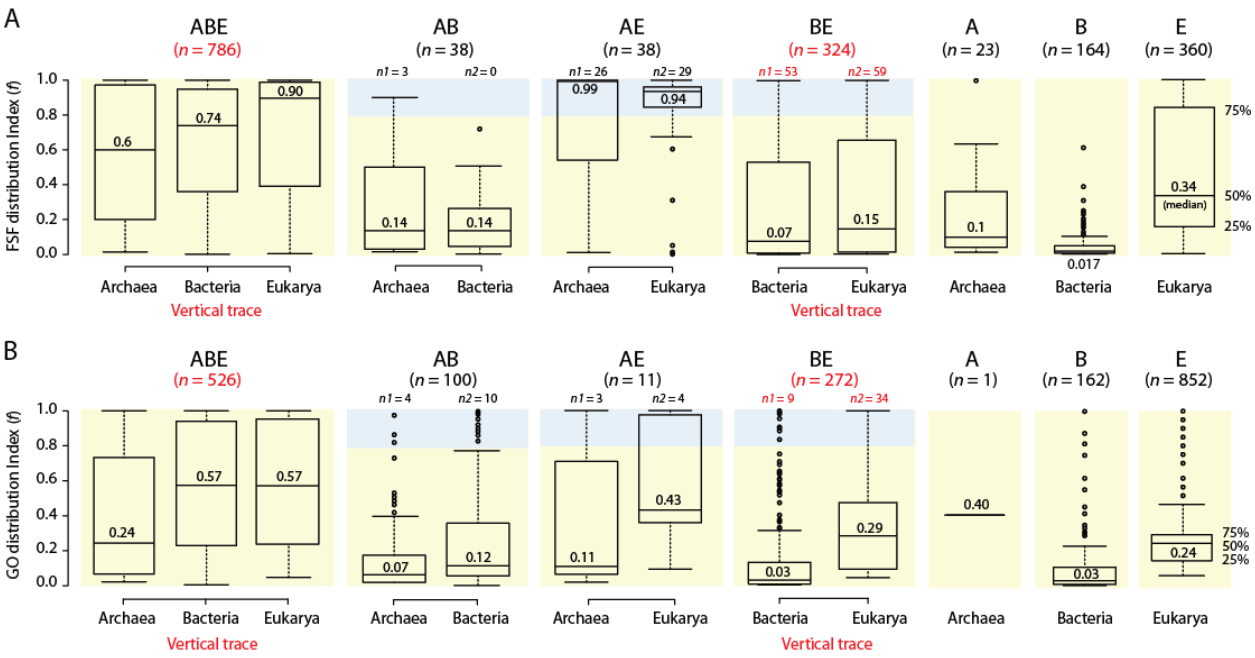

Figure 3.3 Identification of vertical evolutionary trace. The spread of FSF domain structures **(A)** and GO terminal terms **(B)** in the proteomes and functionomes of each member superkingdom in the seven Venn taxonomic groups (panels ABE, AB, AE, BE, A, B and E). Shaded regions indicate FSFs or GOs that were present in >80% of the proteomes ($f > 0.8$), and their numbers, n_1 and n_2. Numbers in boxplots of each distribution indicate group medians. Numbers in red suggest the strongest vertical evolutionary trace.

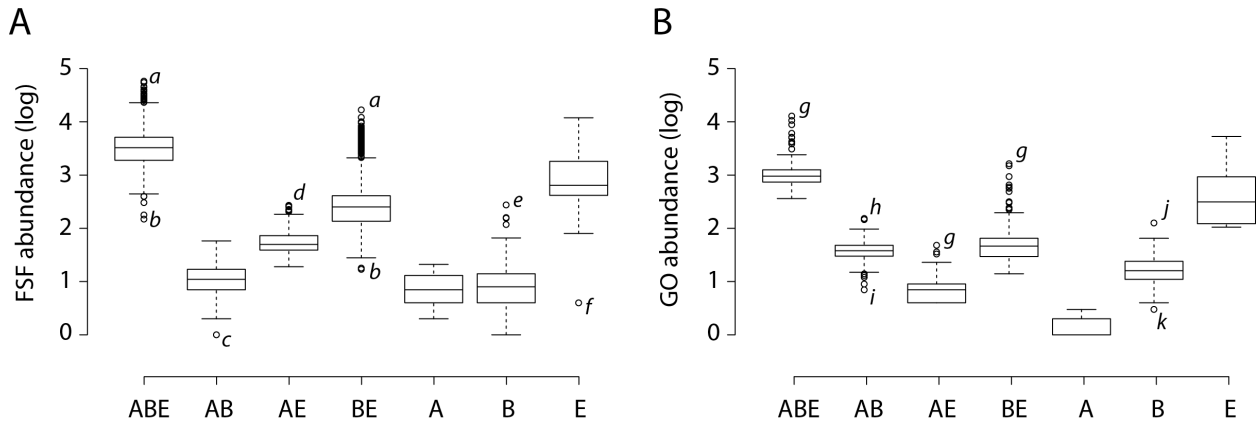

Figure 3.4 Genomic abundance distribution of traits in taxonomic groups. Boxplots comparing the log-transformed abundance values of structural **(A)** and functional **(B)** traits in the proteomes and functionomes of the seven Venn taxonomic groups. Italicized characters identify outliers with maximum and minimum abundance of traits in each group: [a]*Takifugu rubripes;* [b]*Cand.* Hodgkinia cicadicola Dsem*;* [c]*Mycoplasma genitalium G37;* [d]*Zea mays;* [e]*Mycobacterium marinum;* [f]*Guillardia theta;* [g]*Homo sapiens;* [h]*Rhodospirillum rubrum;* [i]*Desulfurococcus kamchatkensis;* [j]*Ralstonia eutropha;* [k]*Thermosipho africanus.*

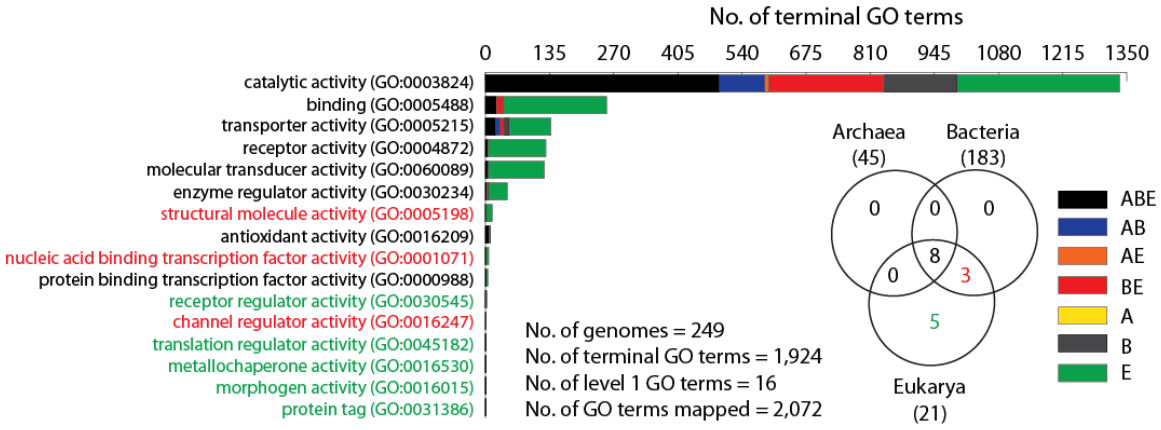

Figure 3.5 Distribution of higher-level molecular functions in taxonomic groups. Barplots illustrating the breakdown of terminal GOs in the seven taxonomic groups for level 1 GO terms. A total of 1,871 out of 1,924 GOs (97.24%) could be reliably mapped to their parents. Level 1 GOs that could not be mapped include 'D-alanyl carrier activity [GO:0036370]', 'electron carrier activity [GO:0009055]', 'chemoattractant activity [GO:0042056]', 'chemorepellent activity [GO:0045499]' and 'nutrient reservoir activity [GO:0045735]'. Note that terminal GOs may have more than one parent. The Venn diagram shows that none of the A, B, AB, and AE taxonomic groups uniquely code for any level 1 GO terms.

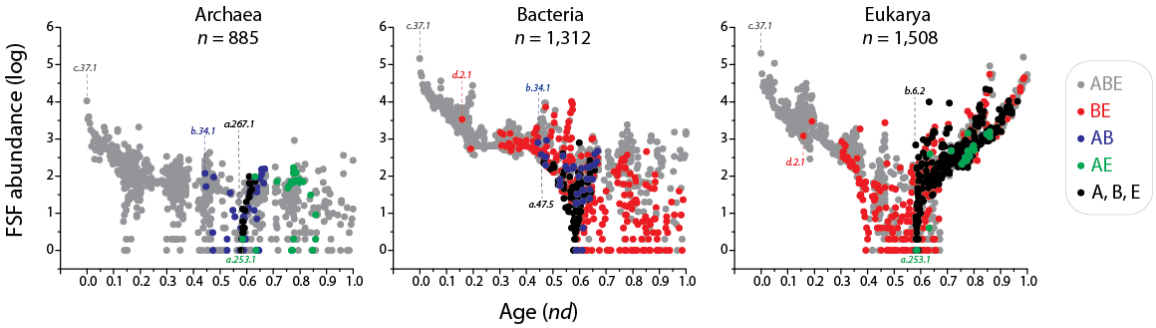

Figure 3.6 Evolutionary timelines highlighting the abundance of FSFs in superkingdom taxonomic groups. Evolutionary age (*nd*) was calculated from a phylogenetic tree of protein domains describing the evolution of 1,733 FSFs (taxa) in 981 organisms (characters) (see [21,38] for technical details). SCOP alphanumeric identifiers were used to identify the most ancient FSF in each taxonomic group. In case of multiple FSFs of same age, only the FSF with maximum abundance was labeled. *c.37.1* is the P-loop containing NTP hydrolase FSF; *b.34.1* is the C-terminal domain of transcriptional repressors FSF; *a.267.1* is the topoisomerase V catalytic domain-like FSF; *a.253.1* is the AF0941-like FSF; *d.2.1* is the Lysozyme-like FSF; *a.47.5* is the FlgN-like FSF; *b.6.2* is the major surface antigen p30, SAG1.

Tables

Table 3.1 List of universal FSFs that were present in all proteomes of the *structure* dataset.

No.	SCOP Id	FSF Id	FSF description
1	52540	c.37.1	P-loop containing nucleoside triphosphate hydrolases
2	50249	b.40.4	Nucleic acid-binding proteins
3	53067	c.55.1	Actin-like ATPase domain
4	51905	c.3.1	FAD/NAD(P)-binding domain
5	53098	c.55.3	Ribonuclease H-like
6	54211	d.14.1	Ribosomal protein S5 domain 2-like
7	55681	d.104.1	Class II aaRS and biotin synthetases
8	50447	b.43.3	Translation proteins
9	54980	d.58.11	EF-G C-terminal domain-like
10	50104	b.34.5	Translation proteins SH3-like domain
11	50465	b.44.1	EF-Tu/eEF-1alpha/eIF2-gamma C-terminal domain
12	55174	d.66.1	Alpha-L RNA-binding motif
13	54768	d.50.1	dsRNA-binding domain-like
14	55257	d.74.3	RBP11-like subunits of RNA polymerase
15	52080	c.12.1	Ribosomal proteins L15p and L18e
16	54686	d.41.4	Ribosomal protein L16p/L10e
17	54843	d.55.1	Ribosomal protein L22

Table 3.2 List of universal GOs that were present in all functionomes of the *function* dataset.

No.	GO Id	GO description
1	GO:0005524	ATP binding
2	GO:0008270	zinc ion binding
3	GO:0000287	magnesium ion binding
4	GO:0005525	GTP binding
5	GO:0004222	metalloendopeptidase activity
6	GO:0010181	FMN binding
7	GO:0030145	manganese ion binding
8	GO:0003924	GTPase activity
9	GO:0003887	DNA-directed DNA polymerase activity
10	GO:0004252	serine-type endopeptidase activity
11	GO:0003746	translation elongation factor activity
12	GO:0009982	pseudouridine synthase activity
13	GO:0004523	ribonuclease H activity
14	GO:0004826	phenylalanine-tRNA ligase activity
15	GO:0004821	histidine-tRNA ligase activity
16	GO:0004820	glycine-tRNA ligase activity
17	GO:0004824	lysine-tRNA ligase activity
18	GO:0004831	tyrosine-tRNA ligase activity
19	GO:0004618	phosphoglycerate kinase activity
20	GO:0004634	phosphopyruvate hydratase activity
21	GO:0004749	ribose phosphate diphosphokinase activity
22	GO:0003952	NAD+ synthase (glutamine-hydrolyzing) activity
23	GO:0004815	aspartate-tRNA ligase activity
24	GO:0004807	triose-phosphate isomerase activity
25	GO:0004813	alanine-tRNA ligase activity
26	GO:0003917	DNA topoisomerase type I activity

Table 3.3 List of FSFs that were uniquely detected in the proteomes of AE taxonomic group.

No.	Scop Id	FSF Id	FSF description
1	48140	a.94.1	Ribosomal protein L19 (L19e)
2	109993	a.222.1	VPS9 domain
3	53032	c.52.2	tRNA-intron endonuclease catalytic domain-like
4	54984	d.58.12	eEF-1beta-like
5	116742	a.60.14	eIF2alpha middle domain-like
6	118310	g.80.1	AN1-like Zinc finger
7	54575	d.29.1	Ribosomal protein L31e
8	55481	d.91.1	N-terminal domain of eukaryotic peptide chain release factor subunit 1, ERF1
9	89124	a.183.1	Nop domain
10	55003	d.58.16	PAP/Archaeal CCA-adding enzyme, C-terminal domain
11	55267	d.75.1	tRNA-intron endonuclease N-terminal domain-like
12	110993	d.58.51	eIF-2-alpha, C-terminal domain
13	69695	d.201.1	SRP19
14	82704	d.68.6	AlbA-like
15	48662	a.137.1	Ribosomal protein L39e
16	56741	e.15.1	Eukaryotic DNA topoisomerase I, N-terminal DNA-binding fragment
17	46950	a.5.6	Double-stranded DNA-binding domain
18	116820	a.4.15	Rps17e-like
19	75689	g.59.1	Zinc-binding domain of translation initiation factor 2 beta
20	143870	d.329.1	PF0523-like
21	88798	d.230.1	N-terminal, heterodimerisation domain of RBP7 (RpoE)
22	140726	a.253.1	AF0941-like
23	89895	d.235.1	FYSH domain
24	144210	g.41.16	Nop10-like SnoRNP
25	47157	a.23.4	Mitochondrial import receptor subunit Tom20
26	52042	c.9.2	Ribosomal protein L32e
27	111278	d.282.1	SSo0622-like
28	75399	d.211.2	Plakin repeat
29	103456	f.23.28	Preprotein translocase SecE subunit
30	63393	g.41.9	RNA polymerase subunits
31	55418	d.86.1	eIF4e-like
32	101576	b.132.1	Supernatant protein factor (SPF), C-terminal domain
33	141562	b.162.1	At5g01610-like
34	55287	d.78.1	RPB5-like RNA polymerase subunit
35	46924	a.4.11	RNA polymerase subunit RPB10
36	109728	a.5.8	Hypothetical protein AF0491, middle domain
37	100966	d.241.1	Translation initiation factor 2 beta, aIF2beta, N-terminal domain
38	88802	d.17.6	Pre-PUA domain

Table 3.4 List of terminal GOs that were uniquely detected in the functionomes of AE taxonomic group.

No.	GO Id	GO description
1	GO:0000213	tRNA-intron endonuclease activity
2	GO:0004579	dolichyl-diphosphooligosaccharide-protein glycotransferase activity
3	GO:0004965	G-protein coupled GABA receptor activity
4	GO:0004164	diphthine synthase activity
5	GO:0030337	DNA polymerase processivity factor activity
6	GO:0017091	AU-rich element binding
7	GO:0030410	nicotianamine synthase activity
8	GO:0004776	succinate-CoA ligase (GDP-forming) activity
9	GO:0008466	glycogenin glucosyltransferase activity
10	GO:0003975	UDP-N-acetylglucosamine-dolichyl-phosphate N-acetylglucosaminephosphotransferase activity
11	GO:0004581	dolichyl-phosphate beta-glucosyltransferase activity

CHAPTER 4: ORIGIN AND EVOLUTION OF THE VIRAL SUPERGROUP[4]

Introduction

Living organisms can be broadly classified into three superkingdoms, Archaea, Bacteria, and Eukarya [107,165], each harboring a fundamentally unique type of cell. Cellular life forms possess several 'hallmark' features that distinguish them from other biological entities, including viruses. For example, cells are metabolically active, possess ribosomes, maintain internal regulation of pH and temperature (homeostasis), and are *always* compartmentalized by lipid membranes. In contrast, viruses lack each of these features, especially ribosomes (*sensu* [199]), and therefore must penetrate the cellular membranes of their hosts and use their metabolic machinery to produce viral progeny. Thus, at first glance, viruses are only infectious to cells. However, the crucial dependency of viral replication in an intracellular environment also creates fertile grounds for genetic innovations [200]. In fact, viruses have been the likely source of many novel genes and machinery for cellular function [201,202]. For example, DNA [203] and the nucleus [204,205] likely evolved in cells to provide short-term selection advantages against invading viruses but eventually served long-term evolutionary goals [70]. Viruses can also transfer genes between species and increase biodiversity [14]. Historically, they have led to important breakthroughs in biology [206] based on key roles in cellular evolution [194,207]. Thus interaction between cells and viruses has likely benefited both entities.

Viral particles (i.e. virions) consist of a replicon (DNA or RNA), protein coat (capsid), and in some cases, lipid envelopes derived from host membranes. Despite being very simple in organization, virions are perhaps the most abundant (especially in oceans [208,209]) and diverse entities on the planet. Already a number of unique virion morphotypes (e.g. droplet, bullet, bottle, and spindle-shaped; see [210-212]) and seven replication strategies have been described in viruses [213]. In comparison, cells only possess dsDNA genomes and synthesize proteins using an RNA intermediate. These arguments, along with the recent discovery of 'giant' viruses that resemble parasitic cells in genome and physical size [214-217], now blur the fine line that

[4]This chapter has been submitted for publication to *PloS Genetics* and is under editorial review.

once separated cells and viruses and prompts revisiting some of the basic concepts related to viral origins, classification, and evolution. These issues are briefly reviewed below.

Different scenarios of viral origin have been proposed (e.g. [194,207,218-222]). The 'virus-first' hypothesis proposes an ancient origin of viruses prior to the ancestors of cells. It is supported by the widespread presence of some key viral proteins (e.g. the 'jelly-roll' fold) in distantly related viruses and their absence in cellular proteomes [207]. The hypothesis however contradicts the current definition of viruses. For example, all extant viruses are dependent upon cells for virion synthesis and enclose their genomes inside elaborate protein capsids. The development of these crucial features necessitates the pre-existence of some kind of cellular structure (i.e. ancient cell) equipped with rudimentary metabolic and translational machinery [194,203]. Moreover, some scientists have convincingly argued that viruses can create novel genes (lacking cellular homologs) during the intracellular stage in their reproduction cycle (e.g. [201,223]) and that cell-like structures appeared very early in evolution (reviewed in [224]). Thus, the idea of a pre-cellular viral world is not widely accepted. In turn, the 'regression' hypothesis considers viruses to be the extremely reduced forms of parasitic cells. The implication is that parasitic microbes undergoing genome reduction may ultimately transform into viruses. However, even the extremely reduced extant bacterial species (e.g. Rickettsia) possess ribosomes and other 'cellular' features that distinguish them from viruses [199]. Again, the scenario may seem more likely if one considers reduction from ancient cells and not modern cells (*sensu* [194]). Another hypothesis considers viruses as autonomous entities that 'escaped' from modern cellular genomes. This scenario fails to explain the presence of unique viral proteins, especially capsids that are believed absent in cells, and is incompatible with genomic data (see Results). However, an escape from ancient cellular genomes may still be more parsimonious than an escape from modern genomes. Interestingly and as convincingly argued [194], the distinction between primitive and modern cells makes it much easier to think about viral origins in the light of the classical virus-first, regression, and escape hypotheses.

More recently, a structural phylogenomic analysis predicted a hybrid hypothesis of viral origin [14,195]. In that study, the origin of large dsDNA viruses was traced back to primitive vesicle-like compartments (i.e. ancient cells) that existed prior to the last common ancestor of modern cells (the redefined 'last universal cellular ancestor'; LUCELLA [195]). We emphasize however that different evolutionary scenarios and their explanatory power must be tested

objectively. Given the massive genetic and phenotypic diversity of the viral world, this task has remained a big challenge.

In terms of classification, the International Committee on Taxonomy of Viruses (ICTV) assigns virus names and puts them into a taxonomic system that closely resembles the classification of cellular organisms [225]. The latest ICTV report (2013) recognizes 7 orders, 103 families, 22 subfamilies, 455 genera, and 2,827 viral species. Under this classification, viral families belonging to the same order are evolutionarily related. However, only 26 viral families have been assigned to an order and the evolutionary relationships of most of them remain unclear. It is expected that the number of unassigned families will continue to increase with the discovery of novel viruses from atypical environments and because genes of many viral families do not exhibit significant sequence similarities [226]. In contrast, the Baltimore classification defines viruses according to their genome type and replication strategy [213]. This method defines seven viral groups (Groups I-VII) that include dsDNA, ssDNA, dsRNA, plus- and minus-ssRNA, and retrotranscribing viruses (ssRNA-RT and dsDNA-RT). However, this approach is not evolutionarily informative as viruses belonging to the same replicon type may or may not have evolved from a common ancestral virus. Another recent proposal is to define novel viral lineages based on the three-dimensional (3D) structural similarities of major viral capsid proteins and virion assembly pathways [227]. Remarkably, it has been observed that viruses infecting very different hosts share strong structural and morphological similarities. Since capsid architectures are believed to be a hallmark of viruses [199,228] and there are very few known capsid protein folds, this approach leads to only few viral lineages and greatly simplifies the overall diversity of viral groups [227]. However, convergent evolution of protein structures and other non-vertical evolutionary processes could weaken the argument. Moreover, the evolutionary role of non-structural viral proteins (e.g. replication enzymes) cannot be completely ignored as well [229]. In addition, resolving 3D structures for capsid proteins in enveloped viruses has remained a challenge [230]. Taken together, there is need for an improved taxonomy of viruses that could meaningfully capture the massive diversity of the virosphere (i.e. the collection of all viruses) and is supported by sound evolutionary data.

The recent discovery of giant viruses [214-217] has led some to argue that viruses, especially those with large genomes, should be included in the ‘Tree of Life’ (ToL) and be recognized as a ‘fourth’ domain of life [14,231-235]. However, their inclusion in the ToL

implies establishing that viruses are indeed living organisms and recognizing that they evolved either from a single ancestral cell or progenitor virus (i.e. monophyletic origin) or appeared multiple times in evolution via different mechanisms (polyphyletic origin). Historically, the idea that viruses are living organisms has been rejected because (I) they lack their own metabolism, and (II) cannot replicate and evolve outside their hosts, two fundamental properties used to define life (discussed in [236,237]). However, counter-arguments have recently gained popularity especially inspired by the study of 'virus factories', intracellular structures formed by many giant viruses inside infected cells [238]. The virus factory is a 'cell-like organism' (*sensu* [239]), which is compartmentalized by a membrane, possesses ribosomes, obtains energy from mitochondria, and contains full information to successfully produce numerous virions [238]. It is strikingly similar to many intracellular parasitic bacteria that also depend upon host metabolism to reproduce. For these reasons, it has been argued that the true 'self' of a virus is the intracellular 'virus factory' of infected cells, which is metabolically active and should be contrasted with the extracellular and metabolically inert virion state. Under this view, virions are functionally analogous to bacterial spores, plant seeds, and human spermatozoa that solely disseminate genetic information but do not represent the true self of the species [239]. Similarly, the 'virocell' concept [240,241] also emphasizes on the intracellular stage of the virus reproduction cycle, when a virus-infected cell transforms into a virocell and produces virions instead of dividing into daughter cells. In other words, it becomes much easier to think of viruses as living organisms when the concept of viruses being virions is replaced by a focus on the intracellular stage of the virus reproduction cycle. The second argument that viruses do not replicate or evolve independent of cells and hence should not be deemed worthy of 'living' status [236] has been toned down since each species replicates and evolves in nature and requires co-existence with other life forms [242]. Moreover, numerous bacterial parasites survive as obligate endosymbionts of other species and are still considered living organisms. In light of these arguments, we contend that it is legitimate to study viral origins and evolution on a scale comparable to that of Archaea, Bacteria, and Eukarya and to ask fundamental questions related to the evolutionary history of cells and viruses.

However, besides problems of interpretation, numerous technical issues also complicate testing hypotheses of viral origin and evolution. First, most viruses are too small to be seen with the light microscope and cannot be cultured in the laboratory. Hence, their diversity can be easily

underestimated. Second, our knowledge about preserved viral fossils is very limited. Perhaps the best-known examples are the endogenous retroviruses that have become part of our germ-line DNA [243] and the very recent discovery of virus-like particles preserved in geothermal systems [244]. Consequently, their long-term evolutionary trajectory can only be inferred from life cycles and the molecular makeup of extant viruses and cells. Third, viruses evolve much faster than cellular organisms (especially RNA viruses [245]). For example, antigenic shift and drift in influenza viruses generates new viral strains every season [246]. High mutation rates make it very difficult to unify viral families using sequence-based phylogenetic analyses, a challenging problem that also plagues the deep evolutionary study of cellular organisms. In fact, homologous proteins often diverge beyond recognition at sequence level, especially if long evolutionary time has passed [247]. In such cases, traditional sequence-based homology searches (e.g. BLAST) and alignment software perform very poorly. However, the 3D packing of amino acid side-chains in the protein domain structure cores retains its arrangement over long evolutionary periods [201]. Because homologous proteins often maintain 3D fold and biochemical properties, they can still be recognized at the structure and function levels [34,248-252]. Therefore, it is crucial to utilize molecular structure (and/or functional) information when studying deep evolutionary patterns such as those involving viruses [201].

One popular scheme for classifying protein domains based on their structural, functional, and evolutionary relationships is the Structural Classification of Proteins (SCOP) database [33,34]. SCOP classifies protein domains of known 3D structure into a hierarchy of fold families (FFs), superfamilies (FSFs), folds, and classes. Protein domains belonging to a common FF typically show high sequence identity (>30%). In turn, sequence identity for FFs grouped together into a common FSF is very low (generally <15%) but there is convincing structural and/or functional evidence that suggests common ancestry of these domains. Folds unite FSFs harboring common structural core topologies, while classes define the type of secondary-structure present in protein domains (i.e. all-α, all-β, α/β, α+β, and others). FSFs (and FFs) provide meaningful evolutionarily information and are generally more conserved in evolution than protein sequence [27,29,201]. This is demonstrated by the fact that nearly half a million-protein sequences in UniProtKB/Swiss-Prot [253] map only to ~1,200 SCOP folds and about ~2,300 FSFs (SCOP 1.75). Empirically, it has been shown that structure is at least 3-10 times more conserved than protein sequence [35]. Moreover, the conserved 3D core of FSF domains

rarely (i.e. 0.4-4%) evolves by convergent evolution [122]. A focus on FSF domains also enables to put bounds on the molecular diversity of viruses and cellular organisms. This and other advantages (see [37]) make FSF domains useful characters to use in evolutionary studies, especially when the focus is to reconstruct the deep evolutionary history of life.

Here, we take a data-driven approach to study the evolution of viruses and cells and test several hypothesis and schemes that have been historically proposed to describe viral evolution and classification. We analyzed 5,080 completely sequenced proteomes of viruses and cells and assigned FSF domains to their proteins using structure-based hidden Markov models (HMMs) defined by the SUPERFAMILY database (ver. 1.75) [40,41]. Applying both comparative genomic and phylogenomic strategies, we ask a number of crucial questions: Can we quantify viral diversity? How many unique protein folds exist in the virosphere? What is the predominant direction of gene transfer (i.e. cell-to-virus or virus-to-cell)? Are viruses infecting different organisms evolutionarily related? Does structure lead to a better taxonomy of viruses? Are viruses monophyletic or polyphyletic? Where do viruses lie on the ToL? And what were the earliest replicons?

Remarkably, we show that despite exhibiting very high levels of molecular diversity, viral proteomes retain traces of their evolutionary history that can be recovered using advanced bioinformatics approaches. The most parsimonious hypothesis inferred from genomic data suggests that modern viruses originated from ancient cells that harbored segmented RNA genomes and co-existed with the ancestors of modern cellular organisms. We refer to the former entity as ‘proto-virocells’ to emphasize the cellular nature of ancient viruses and to make clear their distinction from modern virocells that produce elaborate virions [240,241]. In turn, we use LUCELLA to refer to the last common ancestor of modern cells with understanding that it was not the first cell. This implies the existence of ancient cellular lineages prior to LUCELLA dating back to the ‘last universal common ancestor’ (LUCA) of both cells and viruses. According to our data, the prolonged pressure of genome and particle size reduction eventually reduced proto-virocells into extant viruses (identified by the complete loss of their cellular nature), while other co-existing lineages gradually diversified into modern cells. Interestingly, the cellular nature of viruses is restored today once modern viruses (re)-take control of the cellular machinery of modern cells (i.e. modern day virocells) or when they integrate into cellular genomes.

The new model for the origin and evolution of the viral supergroup captures the many aspects of viral diversity (e.g. host preferences, viral morphologies, proteomic makeup) and is backed by strong support from genomic data. It is also partially compatible with existing models of viral origins as our data confirm an ancient history of the proteomes of viral ancestors (virus-first), evolution by gene loss (regression), and escape from primitive cells.

Methods

Data Retrieval

Viral protein sequences were retrieved from the NCBI Viral Genomes Resource (June 2014) [254]. A total of 190,610 viral proteins corresponded to proteomes of 3,966 viruses. For simplicity, unclassified and unassigned phages and viruses, and deltaviruses that require helper co-viruses to replicate in host tissues (e.g. *Hepatitis delta virus*) were excluded from the analysis. Viral proteomes were scanned against SUPERFAMILY HMMs [40,41] to detect significant SCOP FSF domains (*E*-value < 10^{-4}). Proteomes with no hits were further excluded from the analysis. This yielded a final viral dataset of 3,460 proteomes including 1,649 dsDNA, 534 ssDNA, 166 dsRNA, 991 ssRNA (880 plus-sense and 111 minus-sense), and 120 retrotranscribing (56 ssRNA-RT and 64 dsDNA-RT) viruses. In turn, FSF assignments for 10,930,447 proteins in the proteomes of 1,620 cellular organisms were directly retrieved from the local installation of SUPERFAMILY MySQL database (release July 2014; ver. 1.75). The cellular dataset included 1,620 proteomes from 122 Archaea, 1,115 Bacteria, and 383 Eukarya. A total of 1,995 significant FSF domains were detected in ~11 million proteins of 5,080 proteomes sampled from cells and viruses. We labeled FSF domains by SCOP *concise classification strings* (*css*) for quick identification. For example, the 'P-loop containing NTP hydrolase' FSF is c.37.1, where 'c' is the α/β class of secondary structure present in the protein domain, '37' the fold, and '1' the FSF.

Maximum-Parsimony (MP) Tree Reconstruction

Phylogenomic analysis was carried out as previously described [29,43]. Specifically, we calculated the abundance (i.e. total redundant count) of each FSF in every proteome. Raw abundance values were log-transformed and rescaled to ensure compatibility with PAUP* (ver. 4.0b10) [44]. For example, the raw abundance value of FSF *a* in proteome *b* was log-transformed (g_{ab}) and then divided by the maximum abundance value in that proteome (g_{ab_max}). This was done for each FSF in every proteome. The transformed matrix was then rescaled from 0-23 to yield 24 possible character states for use in PAUP* (see equation below).

$$g_{ab\_normal} = \text{round}\left[\ln(g_{ab} + 1) / \ln(g_{ab\_max} + 1) * 23\right]$$

Normalization and rescaling ensure compatibility with PAUP* and protect against the effects of unequal proteome sizes and variances. MP was used to reconstruct trees of domains (ToDs) and trees of proteomes (ToPs). ToDs describe the evolution of FSF domains (taxa) using proteomes as characters. In turn, ToLs resemble conventional phylogenies that describe the evolution of proteomes (taxa) using FSF domain characters. Trees were rooted by the Lundberg method [47] that does not require specification of any outgroup taxon. Instead, first an unrooted network is calculated which is rooted *a posteriori* by the branch yielding minimum increase in overall tree length. For this purpose, ancestral character states were specified using the ANCSTATES command in PAUP*. ToDs were polarized by the maximum character state, assuming that the more abundant and widespread FSFs should be more ancient relative to those with lower abundance and limited spread. In contrast, ToLs were rooted by the minimum character state assuming that modern proteomes evolved from a relatively simpler urancestral organism that harbored only few FSFs [38]. We note that MP approximates maximum likelihood when phylogenetic characters evolve at different rates [101] and is appropriate for global proteome studies. Bootstrap (BS) analysis with 1,000 replicates was performed to assess the reliability of deep evolutionary relationships. Trees were visualized using Dendroscope (ver. 3.2.8) [49].

Evolutionary Age of FSF Domains

From the ToD, we calculated a *node distance (nd)* value for each FSF taxon. This distance was given on a relative scale from 0 to 1 and was calculated simply by counting the number of nodes from a terminal taxon to the root node. Thus, FSFs closer to the root had lower *nd* values relative to the more derived FSFs. We have previously shown that *nd* is a reliable proxy for the evolutionary age of FSFs and describes a clock-like behavior of FSF evolution that is remarkably consistent with the geological record [51].

Spread of FSFs in Proteomes

To evaluate the spread of FSFs in the proteomes of cells and viruses, we calculated a distribution index, which we term *f*-value. The *f*-value ranges from 0 to 1 and indicates the fraction of genomes that encode a particular FSF. For example, an *f*-value of 0.75 for FSF X in Archaea, 0.82 in Bacteria, and 0.93 in Eukarya means it was detected in 75%, 82%, and 93% of the archaeal, bacterial and eukaryal proteomes, respectively.

ToL Reconstructions from the Numerical Analysis of Domain Age

Principal coordinate analysis (PCO) was performed using Microsoft Excel XLSTAT plugin [255]. For this reconstruction, proteomes were treated as samples and FSFs as variables. Because, proteomes are composed of FSFs of different ages (i.e. *nd* values), we transformed the FSF occurrence matrix into FSF occurrence * (1-*nd*) matrix, making the matrix a multidimensional space of evolutionary age of domains. The 'reverse age' 1-*nd* transformation ensured we did not lose information about FSFs of very ancient origin (e.g. c.37.1 that had an *nd* of 0 and could be confused with FSFs that were absent in a proteome). Similarly, the transformation ensured FSF absences (domains that have not yet materialized) did not contribute age to the multidimensional temporal space. Next, Euclidean distances were calculated that described pairwise dissimilarity among proteomes. The pairwise phylogenetic distance matrix was used to calculate the first three principal coordinates that described maximum variability in data. Effectively, the PCO provided the three most significant loadings that described how component parts (FSFs) contribute to the history of systems (proteomes). The proposed evolutionary PCO (evoPCO) should be considered 'rooted' in time as the multidimensional space was centered on an *nd* parameter that correlates with geological time [51]. For reference, we added the previously reconstructed proteome of LUCELLA [38] as an additional sample.

Network and Neighbor-Joining (NJ) Tree Reconstruction

Phylogenomic networks were generated using the NeighborNet algorithm [129] implemented in SplitsTree package (ver. 4.13.1) [130]. An NJ tree was calculated from the pairwise phylogenetic distance matrix using the 'Phangorn' and 'ape' packages in R ver. 2.15.2. For both reconstructions, FSF occurrence was used to characterize randomly sampled proteomes.

Functional Analysis

Gene ontology (GO) [58,59] enrichment analysis was performed using the domain-centric gene ontology resource [56]. A list of FSFs was provided as input and only the most significant ($FDR < 10^{-3}$) and highly specific 'biological process' GO terms that were enriched in the given set of FSFs were retrieved.

Results and Discussion

The Proteomes of Cells and Viruses Overlap in Genetic Complexity

A Venn diagram shows that roughly two-third of total FSFs (1,279 out of 1,995) were only present in cellular proteomes (Figure 4.1A). But viruses shared FSFs with each and every possible Venn group, indicating evolutionary continuity between cells and viruses. The most popular Venn groups of universal FSFs found in both cells and viruses (ABEV) or shared by Archaea, Bacteria, and Eukarya (ABE) had 442 and 457 FSFs, respectively. The large size of ABEV group possibly suggests a common and ancient origin (co-existence) of viruses with cells, very much like the large size of ABE strengthens the hypothesis of a common ancestor of modern cells (i.e. LUCELLA). However, not all ABEV or ABE FSFs should be considered vertically inherited, since they could be subject to both vertical inheritance and horizontal gene transfers (HGT) [256]. Thus, the proportion of vertically inherited FSFs in higher order Venn groups must be determined (read below).

In turn, FSFs unique to superkingdoms and viruses (i.e. A, B, E, and V groups) indicate novel gains specific to each supergroup. These gains were more common in Eukarya (283 novel FSFs) and Bacteria (154 FSFs) than in Archaea (24 FSFs) and viruses (66 FSFs) (Figure 4.1A). Remarkably, the 66 FSFs unique to viruses (Table 4.1) were ~3 fold greater in number than the corresponding archaeal FSFs. Previously, Abroi and Gough (2011) identified 63 virus-specific FSFs (VSFs) in their analysis of viral proteins from UniprotKB [201]. Among those, 51 were also present in our census along with 15 new VSFs (Table 4.1). However, 12 FSFs from [201] (of which 11 were common to our dataset) were no longer part of the V group. One example is the ‘Group II dsDNA viruses VP’ FSF (b.121.2), which is the ‘double jelly-roll’ capsid fold signature of many dsDNA viruses [226]. We discovered that b.121.2 was completely absent in prokaryotes and was rare in eukaryotes, i.e. detected only in 5 eukaryal proteomes (1.3%). Thus, it was likely transferred horizontally to few eukaryotes from their respective dsDNA viruses. Similarly, the ‘Major capsid protein VP5’ FSF (e.48.1), which includes the capsid protein of herpesviruses, was present only in one eukaryotic proteome (0.3%), suggesting another virus-to-host HGT event. Another example was the ‘Influenza hemagglutinin (stalk)’ FSF (h.3.1), which was categorized in the ABEV group (Table 4.1). However, it was detected only in one archaeal (0.8%), 14 bacterial (1.3%), and 3 eukaryal (0.8%) proteomes, indicating its rare presence in

cells. In turn, and as the name suggests, it was detected mostly in influenza viruses and only in one dsDNA virus (Lactobacillus johnsonii prophage Lj928). These observations suggest that VSFs are spreading to other Venn groups, sometimes involving HGT episodes from viruses to both cells and other viruses. Also the large size of ABEV group is simply not a consequence of gene gain by viruses from their hosts but is bidirectional. Importantly, this predicts that the actual number of VSFs could be under-represented in our census and is expected to grow, once a pool of more diverse viruses is sequenced and HGT-associated relationships are identified. Furthermore, viral genomes can often integrate into cellular genomes and contribute some proteins to their make up. These proteins would be included in other Venn groups such as AB, ABV, BE, and others. Thus, the actual count of viral FSFs may be even higher.

The Unique Identify of the Viral Supergroup

VSFs are hallmarks of viruses. They include proteins involved mainly in viral pathogenesis such as binding to host DNA and receptors, manipulating host immune systems, and encapsulating viral genomes with capsid proteins (Table 4.1 GO [58,59] enrichment analysis confirmed that these proteins establish strong host-parasitic interactions between viruses and cells (Table 4.2). Therefore, VSFs must be linked to the appearance of parasitism in viruses once parasitic lifecycles were established. VSFs challenge the idea that viruses are merely ‘gene robbers’ and capture cellular genes via HGT [237]. In turn, they uniquely identify the viral supergroup on a scale comparable to that of Archaea, Bacteria, and Eukarya, each of which also encodes their set of unique FSFs (Figure 4.1A).

However, the existence of VSFs begs the question about their source of origin. Classical explanations are their gain by HGT from cellular species that are yet to be sequenced or from a yet-to-be-discovered ‘fourth’ domain of life. The former scenario is less likely as the number of VSFs does not decrease with an increase in sequenced genomes (also argued in [223]), while the latter is difficult to prove. In turn, a more parsimonious explanation is the origin of VSFs during the virocell stage [240,241] in virus reproduction cycle when they have full access and control over cellular machinery and can create new genes by different mechanisms such as *de novo* gene creation and gene duplication [223]. This is supported by the fact that although VSFs were detected in all seven viral subgroups, they were mostly specific to them (Table 4.1). Only 3 VSFs were shared by more than one viral subgroup. These included the ‘Coronavirus S2

glycoprotein' FSF (h.3.3) shared by plus-ssRNA (coronaviruses) and dsDNA (*Cafeteria roenbergensis virus*) viruses, the 'Influenza hemagglutinin (stalk)' FSF (h.3.1) shared by minus-ssRNA (influenza viruses) and dsDNA (L. jonhsonii prophage Lj928) viruses, and the 'Viral protein domain' FSF (b.19.1) shared by dsRNA (rotaviruses), plus-ssRNA (coronaviruses) and minus-ssRNA viruses (influenza viruses) (Table 4.1). In the first two cases, the possibility of virus-to-virus HGT from RNA to DNA viruses cannot be ruled out, while b.19.1 could be a unifying feature of most RNA viruses (confirmed below).

The implications of the discovery of VSFs are two fold: (I) they constitute the likely subset of viral proteins that could become hot targets for drug discovery and medical applications, and (II) their mere existence strongly argues that viruses are capable of creating new genes and protein folds.

Viral Proteomes are Enriched with Proteins of 'Unknown' Origin

It is often argued that viral genomes only grow by acquiring genes from their hosts [222,257]. To test if this argument is supported by proteomic data, we classified viruses according to their host type into archaeoviruses, bacterioviruses, and eukaryoviruses and studied their proteomic composition (Figure 4.1B). In all cases, viral proteomes contained three classes of proteins: (I) those for which no structural relative was detected in the HMM library, (II) those for which homologs existed in the cellular proteomes, and (III) proteins encoding VSFs.

Class I proteins with no structural hits represented the majority of viral proteins. Roughly, 80% of prokaryotic and 75% of eukaryoviral proteins did not exhibit any significant similarity with structures encoded by proteins of their hosts (Figure 4.1B). Class I proteins were also abundant in the recently discovered giant pandoravirus for which ~84% of the genes lacked significant similarity in the sequence databases [215]. Even bacterioviruses that frequently mediate genetic exchanges between bacterial species encode roughly 80% of class I proteins (our data and [258]). This seriously negates the idea that viruses only pick genes from their hosts, as there was no trace of such acquisition in their proteomic makeup. One explanation could be the rapid and fast evolution of 'imported' proteins in viruses. However, it is unlikely that rapid mutation will erase structural cores without affecting core function [259]. It is also inconsistent with the presence of class II proteins that surprisingly remained robust to fast evolution within the same viral proteomes. Moreover, synonymous-to-nonsynonymous substitution rates for

'unique' genes in giant DNA viruses did not vary significantly from the mutation rates of vertebrate proteins [260], indicating that selection pressures on viral and cellular proteins are largely similar [259].

Importantly, many class II proteins were likely transferred from viruses to cells and not from cells to viruses. For example, many mitochondrial genes in eukaryotes were likely acquired from proviruses integrated into the mitochondrial ancestor [261] and RNA polymerase genes of dsRNA viruses were transferred to eukaryotes very early in evolution [262]. Mammalian genomes are also enriched with retroviral-like elements [263,264] suggesting that viral genes have invaded our genomes. In fact, any kind of viral nucleic acid can be endogenized and this phenomenon is not restricted to retrotranscribing viruses [265]. Interestingly, syncytin protein that plays a crucial part in mammalian placenta development is encoded by an endogenous retrovirus [266]. This shows that virus-to-host gene transfer is an important force that has shaped our proteomes and that viruses encode considerable amount of genetic novelty, part of which may be transferred to cells. Importantly, it falsifies the idea that viral genomes only evolve by acquiring genes from cells. Given the very large size of class I proteins and the mere existence of VSFs, the more parsimonious explanation is an ancient origin of these proteins in a cellular ancestor (i.e. proto-virocell) that gave rise to modern viral lineages [267]. Alternatively, class I proteins could originate from modern day virocells, or perhaps from both sources throughout the evolutionary time (read below). An analysis comparing the sharing levels of class I proteins across different viral groups could test their ancient vs. recent origin. Nevertheless, our data confirm that a large fraction of viral proteomes is composed of proteins that are 'alien' to cellular organisms. Their origin cannot simply be explained by cell-to-virus HGT. Thus, viral evolution should take into consideration the global nature of viral proteomes and should not be restricted to single gene analyses.

Reductive Evolution Explains Viral Makeup

A comparative genomic analysis of proteomic use, defined by the total number of unique FSFs in a proteome (Figure 4.1C), and reuse, defined by the total number of FSFs (Figure 4.1D) revealed the strong influence of reductive evolution in viral proteomes, especially in dsDNA viruses. We recovered a linear pattern of proteome growth in viruses, Archaea, Bacteria, and Eukarya. Interestingly, giant viruses such as *Megavirus lba, Pandoravirus salinus*, and others

overlapped many parasitic and symbiotic microbial species (mostly Mycoplasma and Proteobacteria) in their genome size (see the shaded regions). This shows that one unifying property for cells and viruses could be their common parasitic lifestyle. To confirm, and as a control, we plotted FSF use and reuse for viruses and only 'free-living' organisms that eliminated the overlap between large dsDNA viruses and microbial parasites (Figure D1). Interestingly, giant viruses were not too far away from archaeal species that have also experienced genome reduction in the past [20,86].

Reductive evolution is a phenomenon commonly invoked to explain the evolution of many cellular species that have become increasingly dependent upon others for survival [21,268,269]. Notable examples include many bacterial species that have transitioned from 'free-living' microbes to eukaryotic organelles such as mitochondria and plastids. In fact, there are many examples of extreme genome reduction in parasitic organisms from all three superkingdoms of life, including *Nanoarchaeum equitans* (obligate endosymbiont of *Ignicoccus*) in Archaea, *Candidatus* Tremblaya princeps and many Rickettsia and Mycoplasma species in Bacteria, and *Giardia lamblia* and a large group of parasitic protists (Apicomplexa) in Eukarya [21,268,270]. In other words, a strong link between parasitism and genome reduction has been confirmed in all three superkingdoms [268]. The unifying feature of all obligate intracellular parasites is their strict dependence on the host for nutrients and the tendency to reduce genomes in an intracellular environment. Our global survey of proteomic repertoires now extends the concept of reductive evolution to the viral supergroup. Bandea [219,220] and Claverie [267] have previously argued that because viruses and microbial intracellular parasites share an obligate parasitic lifestyle, they must also evolve in a similar way (i.e. by streamlining their genomes). It is surprising that this phenomenon has rarely been invoked to explain the limited viral makeup despite the fact that viruses *strictly* adhere to an intracellular parasitic lifestyle. Instead, viral evolution is largely explained by gene uptake via HGT, which only plays a minor role in viral evolution and is not supported by comparative proteomics data (Figure 4.1B). In fact, a large number of viral proteins lack homology to the cellular proteins and the existence of VSFs challenges the notion that viruses are simply 'gene robbers' (as claimed in [237]). Taken together, our results and background knowledge support the central assumption that the proteomes of the viral supergroup, especially giant DNA viruses, have evolved by gradual genetic loss (or by becoming refractory to genetic gains).

Viruses Fuel Cellular Diversity

To infer what was the predominant direction of gene transfer, virus-to-cell or cell-to-virus, we divided FSFs in each superkingdom into two sets: (I) those shared only with cells, and (II) those also shared with viruses (i.e. class II proteins of Figure 4.1B). FSFs specific to each superkingdom (i.e. A, B, and E Venn groups in Figure 4.1A) were excluded as they represent novel gains unique to each superkingdom and *de facto* could not be subject to horizontal transfers (unless they were later completely lost from the donor superkingdom). There were a total of 1,022 FSFs encoded by archaeal proteomes. After excluding 24 Archaea-specific FSFs, 533 (52%) were shared only with Bacteria and Eukarya and 465 (45%) were also shared with viruses. Similarly, of the total 1,535 bacterial FSFs, 154 were Bacteria-specific, 786 (51%) were shared only with Archaea and Eukarya and 595 (39%) were also shared with viruses. Finally, eukaryal proteomes encoded a total of 1,661 FSFs including 283 that were Eukarya-specific, 774 (47%) shared only with other superkingdoms, and 604 (36%) also shared with viruses. Next, we calculated an *f*-value to determine the spread of FSFs in the proteomes of each superkingdom (Figure 4.2). The *f*-value is a proxy for how widespread each FSF is in the modern proteomes and ranges from 0 (complete absence in sampled proteomes) to 1 (universal presence).

In all superkingdoms, FSFs shared with viruses were significantly more widespread in proteomes than those shared only with cells (Figure 4.2A). For example, the median *f*-value in Archaea for FSFs shared only with cells was 0.45 in comparison to 0.59 for FSFs shared with viruses. Similarly, medians increased in Bacteria from 0.30 to 0.62, and most significantly in Eukarya from 0.39 to 0.93. One explanation could be larger sampling of eukaryoviruses relative to archaeoviruses or bacterioviruses in our dataset (2,155 vs. 62 and 1,223). Eukaryoviruses also included recently discovered giant viruses that encode hundreds of proteins [214-217]. However, bacterioviruses on average encode more proteins than eukaryoviruses (Figure 4.1B) but those were not as widespread as in bacterial species. This suggests that massive enrichment of eukaryal species by viral FSFs is a significant outcome and is likely due to Eukarya hosting a large number of viruses from each replicon type relative to Archaea and Bacteria ([212]; also read below). Nevertheless, FSFs shared with viruses were significantly more represented in the individual members of each superkingdom. It is thus likely that viruses mediated the spread of these FSFs by serving as vehicles of gene transfer. It also suggests that viruses are very ancient and most likely infected the last common ancestor of each superkingdom, as viral FSFs were

present in a diverse array of cellular organisms ranging from small microbes to large eukaryotes. Taken together, this data suggests that viruses enhance biodiversity by transferring FSFs within superkingdoms and confirms a similar conclusion derived from an analysis of cells and large-to-medium sized viruses [14].

A breakdown by viral replicon type was also meaningful (Figure 4.2B). In Archaea, nearly all the viral FSFs were well represented in member species. Surprisingly, FSFs shared with RNA viruses were also enriched in archaeal proteomes. Because RNA viruses seemingly cannot carry out a productive infectious lifecycle in Archaea (read below), it is unlikely that they picked these FSFs from archaeal hosts via HGT. In turn, it is more likely that RNA viruses infecting different superkingdoms share FSFs that were retained during their 'de-evolution' from the ancient cells. Similar patterns were also seen in bacterial proteomes (Figure 4.2B). Quite remarkably, FSFs shared with each virus replicon type were almost universal (*f* approaching 1) among the members of the eukaryotic superkingdom. As we will now show, this is consistent with Eukarya hosting a large number of viruses from each replicon type.

Viruses Display Very Narrow Host Ranges

Modern viruses can be unified based on their infectious nature. Interestingly, viruses infecting different hosts share strong structural and morphological similarities [227]. Do they share common protein folds as well? To answer this question, we generated a new Venn diagram describing viral FSF repertoires. FSFs that were shared by archaeoviruses (*a*), bacterioviruses (*b*) and eukaryoviruses (*e*) were pooled into the *abe* Venn group, those shared by viruses infecting different superkingdoms into the *ab*, *ae*, or *be* groups, and those unique to viruses infecting a single superkingdom into *a*, *b*, and *e* groups (Venn group nomenclature avoids ambiguity with that of Figure 4.1) (Figure 4.3A). We stress that FSFs in the *abe* group do not mean these were present in a virus capable of infecting Archaea, Bacteria, and Eukarya (to date no virus is known to infect organisms in more than one superkingdom). Instead, it simply refers to the count of FSFs that were shared between archaeoviruses, bacterioviruses and eukaryoviruses. We discovered that viruses infecting the three superkingdoms shared a total of 68 FSFs (Figure 4.3A, *abe* group). A closer inspection revealed that these FSFs performed crucial metabolic functions and were widespread in cellular proteomes ($f > 0.75$) (Figure 4.3B). Importantly, these FSFs originated very early in evolution (Figure D2, *abe* group) and were detected in a large number of

viruses from each replicon type (Figure 4.3B). In fact, 19 *abe* FSFs (28%) were shared by two or more than two viral subgroups.

It is often argued that because viruses infect all species, they must have originated before modern cells. Here we show that viruses infecting the three superkingdoms possess a very large and conserved structural core that is particularly enriched in crucial metabolic functions believed to be very ancient. This is strong indication of both ancient origin of viruses and their co-existence in the form of ancient cells (proto-virocells). An alternative explanation is the transfer of these FSFs from modern cells to viruses via HGT. However, viruses do not infect hosts separated by large evolutionary distances (i.e. they have very narrow host range [212]). Still these FSFs were detected in seemingly unrelated viruses. Moreover, roughly similar patterns were also observed for the *ab*, *ae*, and *be* FSFs (Figure 4.3). This greatly reduces confidence in cell-to-virus HGT, as the probability of a large number of similar HGT events occurring in very different environments (i.e. different hosts and viruses) is very unlikely.

However, a minor role of HGT cannot be ruled out. In fact, FSFs in *a*, *b*, or *e* Venn groups could be more influenced by HGT, as they represent viruses infecting only a single superkingdom. For example, 5 FSFs that were detected only in archaeoviruses (Figure 4.3A, *a* group), 'Ada DNA repair protein, N-terminal domain (N-Ada 10)' (g.48.1), 'An anticodon-binding domain of class I aminoacyl-tRNA synthetases' (a.97.1), 'Carbamoyl phosphate synthetase, small subunit N-terminal domain' (c.8.3), 'ArfGap/RecO-like zinc finger' (g.45.1), and 'Hypothetical protein D-63' (a.30.5) FSFs (Table D4), appear more 'cellular' than 'viral' in nature. Here, the possibility that archaeoviruses picked these FSFs from archaeal hosts during infection cannot be ruled out with confidence. Interestingly, these FSFs were however more widespread in bacterial and eukaryal proteomes than archaeal proteomes but were absent from their respective viruses (Figure 4.3B). This could be a result of loss of viral lineages from Bacteria and Eukarya, or from reductive evolution in Archaea itself [20,86], which would again negate HGT. In turn, *b* and *e* FSFs were more represented in bacterial and eukaryal proteomes respectively (as expected) and did not have very high *f*-values (Figure 4.3B). Specifically, most of the 198 FSFs unique to bacterioviruses could be a result of HGT from Bacteria to viruses, especially since bacterioviruses are known to mediate gene exchange between bacterial species and most of these FSFs originated very late in evolution (Figure D2, *b* group). However, they only constitute a tiny fraction of the proteomes of bacterioviruses (recall the relatively much

bigger size of class I proteins in bacterioviruses that did not originate in Bacteria). Similar patterns were also observed for *e* FSFs (Figure D2, *e* group). Thus, HGT again appeared to play a minor role in the evolution of viruses. Finally, we note that only two FSFs were shared by archaeoviruses and eukaryoviruses (*ae*). This is in line with previous understanding that eukaryoviruses are very distinct from archaeoviruses (discussed in [70]). Remarkably, patterns of FSF sharing and distribution of viral counts in hosts are compatible with a root of the ToL in the archaeal superkingdom (*see* Figure 11 in [271]).

In summary, evolution of viruses follows a bidirectional route influenced by both the vertical inheritance of a structural core present in many distantly related viruses (i.e. those infecting more than one superkingdom) and by HGT of new FSFs from modern cells. The common core includes proteins mainly of cellular origin. Some of these are likely remnants of reductive evolution from ancient cells that existed prior to LUCELLA, while others could be products of HGT from hosts.

Capsid/Coat Structure-based Viral Lineages: A New Taxonomy for Viruses?

Viruses infecting different organisms often use conserved 3D protein folds to produce elaborate capsids and show striking similarities in their virion architecture. These observations have led to the proposal of a new structure-based viral taxonomy [227]. Currently, four major viral lineages have been defined for icosahedral viruses (the most commonly seen capsid symmetry). These include the 'picornavirus-like lineage', 'PRD1/Adenovirus lineage', 'HK97-like lineage' and 'BTV-like lineage' [227]. These lineages capture many viral families and attempt to simplify the overall diversity of the virosphere. The implication is that viruses belonging to one lineage may have a common origin. However, different lineages are not necessarily monophyletic [230].

To test this taxonomy and to determine how the signature FSFs of each lineage distributed in our dataset, we scanned viral proteins against the library of structure-based HMMs. To reduce any false-positives, we used a very conservative *E*-value ($< 10^{-4}$) when assigning FSF domains to viral proteins (see Methods). This likely resulted in missing some hits to known viral protein domains but also protected from unreliable assignments. Using a keyword search on 'capsid' and 'coat' in SCOP 1.75, we identified 20 capsid/coat-related FSFs involved in the assembly and building of viral capsids. Additionally, 6 more FSFs were identified from the

literature. Of those, 22 were detected in the proteomes of sampled viruses (Table 4.3) and were used to classify viruses into structure-based viral lineages. Results were benchmarked against previous knowledge [227].

(1) *Picornavirus-like lineage:* This lineage is characterized by the 'jelly-roll' or 'β-barrel' fold, which is commonly seen in RNA viruses. It is the largest viral lineage, currently including members from plus-ssRNA (*Bromoviridae*, *Caliciviridae*, *Comoviridae*, *Dicistroviridae*, *Luteoviridae*, *Nodaviridae*, *Picornaviridae*, *Sequiviridae*, *Tetraviridae*, *Tombusviridae*, *Tymoviridae*), dsRNA (*Birnaviridae*), ssDNA (*Microviridae*, *Parvoviridae*), and dsDNA (*Papillomaviridae,* and *Polyomaviridae*) viruses but no minus-ssRNA and retrotranscribing viruses [227]. The 'jelly-roll' fold has a topology of eight β-strands organized into two antiparallel sheets and is represented by the 'Nucleoplasmin-like VP (viral coat and capsid proteins)' SCOP fold (b.121), which includes seven FSFs: (I) 'PHM/PNGase F' FSF (b.121.1) involved in oxidation-reduction metabolic processes (not detected in any viral proteome), (II) 'Group II dsDNA viruses VP' FSF (b.121.2), which is the 'double β-barrel' fold signature of the PRD1/Adenovirus-like lineage (read below), (III) 'Nucleoplasmin-like core domain' FSF (b.121.3) involved in the assembly of nucleosomes in cells, and (IV-VII) FSFs b.121.4, b.121.5, b.121.6, and b.121.7 that were detected in the members of the picornavirus-like lineage (read below).

'Positive stranded ssRNA viruses' FSF (b.121.4) was detected in most RNA viruses including plus-ssRNA (14 families), dsRNA (*Birnaviridae*) and also minus-ssRNA (*Lettuce ring necrosis virus*) viruses and defines an important 'Ariadne's thread' (read below). Thus, our computational approach extended the picornavirus-like lineage to also include minus-ssRNA viruses. Experimental work is required to confirm if these viruses truly belong to this lineage. 'ssDNA viruses' FSF (b.121.5) was detected in many ssDNA viruses of the *Microviridae* and *Parvoviridae* families. The capsid and spike proteins (F and G) of *Bacteriophage phiX174* (*Microviridae*) possess the same 'jelly-roll' fold [272] and were reliably matched to b.121.5. Similarly, 'Group I dsDNA viruses' FSF (b.121.6) included coat and L1 proteins from polyomaviruses and papillomaviruses, both established members of the picornavirus-like lineage. Another novel addition was the 'Satellite viruses' FSF (b.121.7) that was detected in the *Circovirus-like genome RW_B* virus (ssDNA). It seems that the coat protein of this virus resembles the 'jelly-roll' coat proteins of satellite viruses that were excluded from our analysis.

The coat protein of satellite viruses (e.g. *Satellite panicum mosaic virus*) harbors a typical 'jelly-roll' fold but can have 1-2 additional β-strands [273]. Thus, this FSF could be another specialized form of the 'jelly-roll' fold.

In our opinion, FSFs b.121.4, b.121.5, b.121.6, and (possibly) b.121.7 could be used to recruit new members of the picornavirus-like lineage. The other members of the b.121 fold either include proteins specific to cells (i.e. b.121.1 and b.121.3) or advanced forms of the 'jelly-roll' (b.121.2) that make a lineage of their own (read below). Importantly, this lineage now includes viruses with all replicon types except the two groups of retrotranscribing viruses and supports the idea that viruses with different replicons may share strong structural and molecular properties. The exercise also revealed that structural relatives of the 'jelly-roll' fold are found in cells (e.g. histone chaperones and metabolic folds) [274-276] and thus it may not be a unique virus hallmark.

(2) *PRD1/Adenovirus lineage:* This lineage includes dsDNA viruses that infect the three superkingdoms. The prototype members include the human adenoviruses (*Adenoviridae*), *Paramecium bursaria chlorella* viruses (*Phycodnaviridae*), the *Bacteriophage PRD1* (*Tectiviridae*), and archaeal *Sulfolobus turreted icosahedral virus* (*Turriviridae*). The lineage is characterized by the 'double jelly-roll' fold, which likely formed by the duplication of the 'jelly-roll' fold [226]. However, the 'jelly-roll' and 'double jelly-roll' folds are utilized differently in assembling capsids and hence form two distinct lineages [226]. Capsids of viruses belonging to PRD1/Adenovirus lineage are assembled in trimers consisting of two β-barrels arranged around a pseudo six-fold axis. The 'double β-barrel' fold corresponded to 'Group II dsDNA viruses VP' FSF (b.121.2) and was detected in *Adenoviridae*, *Ascoviridae*, *Asfarviridae*, *Iridoviridae*, *Mimiviridae*, *Phycodnaviridae*, and *Tectiviridae*. Notable exceptions from [227] were of *Poxviridae* and *Corticoviridae*. However, the 'double β-barrel' protein domain in poxviruses only facilitates virion formation and does not become part of the capsid [207]. New additions were of *Ascoviridae*, *Asfarviridae*, and *Mimiviridae* that were confirmed in another study [226]. The 'double β-barrel' is apparently a virus hallmark and was detected in only 5 out of the 1,620 cellular proteomes (Table 4.3), suggesting it was likely acquired in the few cellular proteomes from their viruses by HGT.

(3) *HK97-like lineage:* This lineage includes tailed viruses belonging to archaeal and bacterial *Caudovirales* (*Myoviridae*, *Podoviridae*, and *Siphoviridae*) and the eukaryotic *Herpesviridae*. The prokaryotic members of HK97-like lineage are highly abundant in oceans and play important roles in regulating the ecosystems. These viruses are also successful pathogens of both prokaryotes and eukaryotes [212]. In our dataset, the HK97 fold corresponded to two FSFs. the 'Major capsid protein gp5' (d.183.1) from *Bacteriophage HK97* and 'Major capsid protein VP5' (e.48.1) from *Herpes simplex virus 1* (Table 4.3). It has been experimentally verified that the 'floor' domain of herpesvirus VP5 and HK97 gp5 have similar structural organization and are evolutionarily related [277]. Moreover, a small tail similar to that of *Podoviridae* has been detected in the herpesvirus capsid, further supporting their inclusion in the HK97-like lineage [278]. There were no additional SCOP relatives of either d.183.1 or e.48.1. It is the second-lineage after PRD1/Adenovirus lineage that includes viral members infecting the three cellular superkingdoms.

(4) *BTV-like lineage:* The BTV-like lineage currently includes three families of dsRNA viruses, *Cystoviridae*, *Reoviridae*, and *Totiviridae*. Members of these families encode both an outer and inner capsid core. The inner core is evolutionarily conserved and is required within the host cell to avoid apoptotic response against foreign dsRNA genomes [279]. The major core protein VP3, which forms the inner shell of the *Bluetongue virus* capsid, characterizes this lineage. About 120 monomers of VP3 are packed with icosahedral symmetry following a rather unique pattern of subunit assembly [279]. This arrangement was also detected in the *Saccharomyces cerevisiae virus L-A* (*Totiviridae*) [280] and *Pseudomonas phage phi 6* (*Cystoviridae*) viruses [281] suggesting the architecture may be unique to dsRNA viruses [227]. VP3 is a multidomain protein containing three domains that belong to different SCOP FSFs. We discovered that 'A virus capsid protein alpha-helical domain' (a.115.1), 'Reovirus inner layer core protein p3' (e.28.1) and 'L-A virus major coat protein' (e.42.1) FSFs likely corresponded to VP3-like architectures, while the 'Outer capsid protein sigma 3' FSF (d.196.1) was associated with the outer core of the *Reoviridae* capsid. These FSFs were detected in the members of *Reoviridae* and *Totiviridae* (but not *Cystoviridae*). Birnaviruses, which also encode a dsRNA genome, were classified in the picornavirus-like lineage because current knowledge dictates that they exhibit stronger affinity with the 'jelly-roll' fold harboring viruses [230]. Consistent with

the signature folds of PRD1/Adenovirus and HK97-like lineages, none of the three FSFs (a.115.1, e.28.1, and e.42.1) had more SCOP relatives.

(5) *More lineages?* Interestingly, ssRNA-RT (*Retroviridae*) and dsDNA-RT (*Caulimoviridae* and *Hepadnaviridae*) harboring retrotranscribing viruses were not part of any of the four lineages in either [227] or our assignments. Retrotranscribing viruses are typically enveloped and their proteins are difficult to crystalize for structural studies. The capsid protein fold from *Retroviridae* contains an N-terminal domain (5-helix bundle) involved in core formation and a C-terminal domain (4-helix bundle) involved in capsid dimerization [282,283]. These domains corresponded to the 'Retrovirus capsid protein, N-terminal core domain' (a.73.1) and the 'Retrovirus capsid dimerization domain-like' (a.28.3) FSFs and were detected in many viruses belonging to *Retroviridae* (e.g. *Human Immunodeficiency virus-1*). In contrast, the capsid fold from *Hepadnaviridae* (e.g. *Hepatitis B virus*) is also helical (5-helices) and obeys a T = 4 icosahedral symmetry. This fold corresponded to the 'Hepatitis B viral capsid (hbcag)' FSF (a.62.1) and was detected in members of *Hepadnaviridae*.

It has been hypothesized that the C-terminal domain of HIV-1 capsid protein shows significant similarities to the HBV capsid protein suggesting that the two lineages could be evolutionarily related [284]. We note that the capsid fold of *Hepadnaviridae* is arranged in an array-like structure where two long helices form a hairpin that dimerizes into a 4-helical bundle closely resembling the 4-helical bundle of *Retroviridae* capsid a.28.3. However, retroviral FSFs (a.28.3 and a.73.1) did not group with the capsid FSF from *Hepadnaviridae* (a.62.1) according to SCOP classification. Subsequent search against the DALI server [285] also failed to detect any apparent structural homology between the two domains (see [286]). Therefore, more work is required to establish if the capsids from retrotranscribing viruses form more independent lineages or just one. However, capsids from both *Retroviridae* and *Hepadnaviridae* are helical and this is in sharp contrast to the β-sheet rich capsids typically found in other lineages. Similarly, other enveloped viruses (e.g. *Flaviviridae*) are hard to classify based on core capsid proteins. There is indication that instead of the nucleocapsid core, the surface glycoproteins involved in membrane fusion may be more similar to other enveloped viruses [230].

Our computational approach enabled a quick scan of thousands of viral proteins against structure libraries and recovered the previously defined four major capsid-based viral lineages

along with proposals for new additions. Only very few members were missing. This could be a result of using a stringent criterion in assigning FSFs to viral proteins. Importantly, results show that viruses with different replicons and proteome histories could have capsids that are structurally very similar and that HMM-based assignment reproduced the well-known viral lineages. We note however that morphological similarities in viruses could also result from convergent evolution, especially because there are only a limited number of 'economical' ways to pack viral genomes. Thus, it is important to consider both the structural (capsid) and non-structural (polymerases and hydrolases) proteins when studying viral evolution. Another obvious shortcoming is the lack of classification for enveloped viruses. We therefore conclude that while the proposal of capsid structure-based viral classification seems promising, more work is required to establish boundaries within the virosphere. Remarkably, the HMM-based computational exercise impressively complements the experimental-based research.

Do Cellular Proteomes Encode Viral Capsid Homologs or Capsid-like Architectures?

Typically, sequence-based approaches have failed to detect counterparts of viral capsids and coat proteins in cellular proteomes. To confirm if indeed capsid/coat related FSFs were exclusive of viruses, we checked for the presence of 22 capsid/coat related viral FSFs in the 1,620 cellular proteomes that were sampled. Out of the total 22 FSFs, 19 were either completely or near-completely absent in cells (Table 4.3). This shows that structural relatives of very few viral capsids exist and thus are extremely rare in the cellular world. Only the 'Major capsid protein gp5' FSF (d.183.1) of *Caudovirales* (HK97-like lineage) was present in ~24% of the cellular proteomes. Interestingly, the HK97-like fold has been detected in the shell-forming protein (encapsulin) of some archaeal nanocompartments that store metabolic enzymes [287]. These nanocompartments are polyhedral protein shells that are morphologically similar to icosahedral viruses. Because archaeal and bacterial encapsulins are homologous, it is likely that prokaryotic protein compartments are closely related to ancient viral capsids [223]. Another example is of bacterial carboxysomes that are also morphologically similar to viral capsids [288] but are built from protein folds not yet detected in viruses [289]. To confirm, we identified two FSFs that are part of bacterial carboxysomes, (I) 'Ccmk-like' (d.58.56), and (II) 'EutN/CcmL-like' (b.40.15) FSFs. Both had an *f*-value of 0 in sampled viral proteomes confirming a lack of overlap between carboxysomes and viral capsids. However, this could in fact represent a loss of an ancient capsid protein fold from modern viruses or could be an outcome of sampling biases

[223]. It is possible that viruses harboring similar folds exist in nature but remain to be discovered. An interesting analogy could also be made for eukaryotes where histone monomers assemble around DNA to produce chromatin structure. Remarkably, this process is mediated by histone chaperones that harbor the 'jelly-roll' fold [274] that is so abundant in icosahedral viruses. Thus, capsid folds and capsid-like architectures may not be unique to viruses. Interestingly, viral capsids store nucleic acids whereas prokaryotic compartments (carboxysomes and encapsulin protein shells) store enzymes. Perhaps the switch from storing proteins to storing nucleic acids facilitated viral origins in an ancient cell [287]. A corollary is the existence of on overlap between the protein shells of viruses and prokaryotic microbes that has been confirmed (at least for encapsulin proteins). Thus, based on current knowledge, although most viral capsid/coat FSFs have no SCOP structural relatives and lack cellular homologs (Table 4.3), rare capsid structural homologies in cellular proteomes suggest either instances of virus-to-host HGT or relics of ancient coexistence of cells and viruses. These findings question the concept of capsids being true virus hallmarks [199,228].

Ariadne's Threads Point to the Early Origin of Viral RNA Replicons

We explored how the 716 viral FSFs distributed between viral replicon types (Figure 4.4). The majority of viral FSFs were only detected in dsDNA viruses (Figure 4.4A). In comparison, proteomes of the ssDNA, ssRNA, dsRNA, and retrotranscribing groups were genetically poor. The dsDNA viruses were also the most represented in our dataset and encoded more proteins than ssDNA and RNA viruses (Table 4.4). Roughly, 91% (649 out of 716) of the total viral FSFs were unique to a single viral subgroup and only ~9% (67) were shared by more than one subgroup (Figure 4.4A). Generally, the number of shared FSFs in each viral subgroup exceeded the number of unique FSFs except for dsDNA and minus-ssRNA viruses. The substantial number of 586 unique FSFs in dsDNA viruses is especially noteworthy. One explanation could be the very large number of dsDNA viruses sampled in our study relative to other subgroups. However, the proteomic coverage (i.e. number of FSFs in a proteome / total number of proteins) of dsDNA viruses was only 26% and was second lowest in our dataset (Table 4.4). A better explanation is the very large size of proteomes in dsDNA viruses that was adequately translated into the size of their FSF repertoires (Table 4.4).

A 7-set Venn diagram made clear that each viral subgroup shared FSFs with every other subgroup (the sole exception being ssDNA and dsDNA-RT viruses), but did so sparsely (Figure 4.4A, Venn diagram). The diagram shows there was no single FSF common to all viral subgroups (Figure 4.4A). However, it also revealed that the minus-ssRNA and dsDNA groups circumscribed the most widely shared FSFs (traces highlighted in the Venn diagram) (Table 4.5). The 'DNA/RNA polymerases' FSF (e.8.1), which includes T7 RNA polymerase, RNA-dependent-RNA-polymerase of plus-sense and dsRNA viruses, reverse transcriptase, DNA polymerase I, and the catalytic domain of Y-family DNA polymerase, was detected in six out of the seven subgroups. Polymerases are crucial for the successful transfer of genetic information to the progeny and were present in all except ssDNA viruses, which replicate by converting into an intermediate double-stranded form using polymerase enzymes from the host. Therefore, the absence of polymerase structures in ssDNA viruses is not surprising.

In turn, two FSFs were detected in five out of the seven viral subgroups. These included the 'P-loop containing NTP hydrolase' FSF (c.37.1) and the 'S-adenosyl-L-methionine-dependent methyltransferases' FSF (c.66.1) (Table 4.5), two of the most abundant and widespread metabolic FSFs in modern cells. Both FSFs were present in all subgroups except retrotranscribing viruses (Table 4.5). An additional two FSFs, the 'Ribonuclease H-like' FSF (c.55.3) and the 'Positive stranded ssRNA viruses' FSF (b.121.4) were detected in four out of the seven viral subgroups (Table 4.5). The c.55.3 superfamily includes many proteins involved in informational processes (including replication and translation) that are universal among cellular proteomes. This FSF was relatively widespread in viral subgroups but was absent in the proteomes of plus-ssRNA, dsRNA and dsDNA-RT viruses. It was especially abundant in the ssRNA-RT (79% of the proteomes) and dsDNA (58%) viruses. The c.55.3 FSF also includes the catalytic domain of retroviral integrase, which is an important target to silence retroviral gene expression [290] and is medically important in treating HIV infections. In turn, b.121.4 is the 'jelly-roll' fold, which is one of the most common topologies observed in viral capsid proteins [226,291]. Finally, 10 FSFs were present in three out of the six viral subgroups, while 52 were shared by two subgroups (Figure 4.4A, Venn diagram; Table 4.5).

Since Venn diagrams of proteomes contain in themselves information about their origin and evolution [256], we applied Ariadne's thread logic to dissect possible vertical evolutionary traces in FSF sharing (Figure 4.4B). We define our Ariadne's threads as Venn subgroups of FSFs

shared by 2-6 of the 7 viral replicon types (there were no FSFs shared by all 7 viral groups). These threads revealed that only 18 out of the 120 possible Venn subgroups of shared FSFs existed (total Venn-internal groups 2^7-1=127), 14 shared by 2-3 viral groups. They make explicit how sparsely shared are FSFs in viral groups and uncover deep evolutionary patterns likely left by reductive evolutionary loss. Only 8 out of 21 and 6 out of 35 possible subgroups shared by 2 and 3 viral groups, respectively, were present. Remarkably, dsDNA viruses, which hold the largest proteomes and comparatively are minimally affected by reductive evolution, were part of 11 out of these 14 Venn subgroups. A total of 9 (64%) of these 14 subgroups with their 39 FSFs (63%) involved minus-ssRNA, plus-ssRNA and dsRNA replicons suggesting a possible viral origin in RNA genomes. Out of 64 possible groups sharing 4-7 replicon types, only 4 groups were present (lines in Ariadne's thread diagram of Figure 4.4B), all of which heavily support a common origin in minus-ssRNA, plus-ssRNA and dsRNA replicons. As mentioned above, these four groups represent polymerases, metabolic enzymes, ribonuclease and capsid-associated FSFs. Finally, a large number of FSFs were shared between DNA and RNA viruses (Figure 4.4C) suggesting that the virosphere may not be as disjoint as previously thought.

In summary, the patchy distribution of FSFs within the viral supergroup revealed a significant overlap between viruses of different replicon types. While the majority of FSFs were unique to a particular subgroup, a large number of FSFs were shared between viruses belonging to different replicon types (Figure 4.4). Most of the central proteins that are involved in key cellular processes were also widespread among viruses further supporting their ancient coexistence with cells.

Reconstruction of the History of FSF Domains

Comparative genomics and Ariadne's threads suggested an early 'cell-like' existence of viruses. These encouraging results prompted a careful phylogenomic analysis of the genomic census of FSF structures in sampled proteomes. The reconstruction of a phylogenomic ToD describing the evolution of 1,995 FSF domains (taxa) in 5,080 sampled proteomes (characters) (see Methods for tree reconstruction protocol) showed that most viral FSFs originated very early in evolution (see the legend bar on top of the ToD in Figure 4.5A). Due to their highly unbalanced nature, ToDs enabled calculation of a 'proxy' for the relative age of each FSF domain, defined as the *nd* value. The *nd* is a relative phylogenetic distance on a scale from 0

(most ancient) to 1 (most recent) and was calculated simply by counting the number of nodes from a terminal taxon to the root node (see [20] for details; also see Methods). To uncover likely evolutionary scenarios, we plotted FSFs in each of the 15 Venn groups of Figure 4.1A against their FSF ages (i.e. *nd* values) (boxplots in Figure 4.5A).

The ABEV Venn group, which includes 442 FSFs encoded by both cells and viruses, was the most ancient group and covered the entire *nd* axis. The 'P-loop containing NTP hydrolase' (c.37.1) was the first FSF to appear at *nd* = 0. The median *nd* was ~0.4, suggesting that at least 50% of the ABEV FSFs originated very early in evolution and were also shared with viruses. This finding is remarkable and implies that some of the most ancient FSFs found in cells were also shared by very different groups of viruses, suggesting again the ancient coexistence of cells and viruses in the form of primitive cells. In turn, the relatively longer tail on the right likely includes many FSFs of recent origin (*nd* > 0.63) that could have been gained in viruses from cells by HGT. The ABEV group was followed by the appearance of the ABE group. The first ABE FSF was the 'ACT-like' FSF (d.58.18), which includes regulatory protein domains mainly involved in amino acid metabolism and transport. We propose that d.58.18 was most likely 'lost' from viruses, as simultaneous gain in three superkingdoms is less likely compared to loss in just one. By extension, the appearance of the BEV group with the inception of the 'Lysozyme-like' FSF (d.2.1) at *nd* = 0.15 signals the loss of first FSF in a cellular superkingdom (Archaea). Simply, absence of an ancient FSF in one group (out of three or four) is more likely a result of reductive evolution than separate gains (as previously described [20]). The previously reconstructed proteome of LUCELLA [38] was reported to encode a minimum of 70 FSFs. The most recent of those FSFs was 'Terpenoid synthases' FSF (a.128.1) that appeared at *nd* = 0.19 and was absent from viruses. These events demonstrate the early reductive tendencies in early cellular lineages, especially in the protocells leading to viruses and Archaea.

In comparison, FSFs unique to superkingdoms and the viral supergroup appeared much later (note the appearances of the A, B, E, and V groups in Figure 4.5A). As explained above, these gains were restricted to only one 'superlineage' and signaled the diversification of that superkingdom or supergroup. The late appearance of VSFs (V group in Figure 4.5A) is interesting as it includes FSFs involved in viral pathogenicity (Tables 4.1 and 4.2). The phylogenomic analysis shows that VSFs originated at the same time or after the diversification of modern cells. Thus, they represent the time point when proto-virocells under prolonged genome

reduction pressure completely lost their cellular nature and became fully dependent on emerging archaeal, bacterial and eukaryal cells for reproduction. In other words, modern virocell lifecycles established once diversified modern cellular lineages appeared in evolution. This idea is strengthened by the evolutionary appearances of the AV, BV, and EV groups soon after that of superkingdom-specific A, B, and E groups, respectively (see the patterned regions in Figure 4.5A). We speculate that FSFs in the AV, BV, and EV groups either perform functions required by viruses to successfully infect their hosts or were simply HGT gains from their hosts (once the modern viral mode of life established). However, a GO enrichment test on EV FSFs showed that these were enriched in biological processes crucial for cellular development and regulation, such as GO:0048483 [autonomic nervous system development], GO:0002062 [chondrocyte differentiation], and GO:0050921 [positive regulation of chemotaxis] (Table 4.6). It is possible that this repertoire was provided to eukaryotes from viruses or was simply gained from their eukaryotic hosts via HGT. In turn, none of the biological processes were enriched in either the AV or BV groups, suggesting HGT may be at play.

Next, we divided viral FSFs into four subgroups: (I) those shared between prokaryotic and eukaryotic viruses (i.e. the *abe* core of Figure 4.3A), (II) other viral-FSFs shared with cells (cyan circles), (III) VSFs (green circles), and (IV) FSFs not detected in viral proteomes (black circles) (Figure 4.5B). Generally, FSFs of the *abe* core were present in greater number of viral proteomes (higher *f*-values) and in more replicon types (Figure D3). Some of the most popular FSFs again included the 'P-loop containing NTP hydrolase' (c.37.1), 'DNA/RNA polymerases' (e.8.1), and 'Ribonuclease H-like' (c.55.3) FSFs. In turn, FSFs shared with cells were relatively less widespread. However, the 'Lysozyme-like' FSF (d.2.1) was detected in a large number of viruses (18%), mostly bacterioviruses. Lysozymes can penetrate bacterial peptidoglycan layers and facilitate viral entry. We speculate that this capability was also transferred to eukaryotic cells from viruses to block bacterial infections in eukaryotes. Another relatively widespread FSF was the 'Origin of replication-binding domain, RBD-like' (d.89.1) that was detected in ~16% of the sampled viruses. Both the *abe* core and FSFs shared with cells spanned the entire *nd* axis. Thus, viral proteomes encode both the very ancient and very derived FSFs. The former group was most likely inherited vertically from the common ancestor of cells and viruses (i.e. LUCA), while the latter could be a result of recent HGT gains from cells or shared innovation. The enrichment of very ancient FSFs in the *abe* core present in viruses infecting the three superkingdoms provides

strong support to their ancient origin. The origin of VSFs, on the other hand, marks the onset of modern virocell lifecycles. Results therefore highlight two important phases in viral evolution: (I) an early cell-like existence of viruses as proto-virocells (the precursors of modern virocells), and (II) a late transition to the viral mode, as we know it today.

Finally, we zoomed into the ABEV group and separated FSFs belonging to each of the seven viral replicon types (Figure 4.5C). In all viruses, regardless of the replicon type, median *nd* values were very low (see white circles) indicating they shared ancient FSFs with cells. Likewise, each viral subgroup had a longer tail towards the right suggesting that HGT may have played evolutionary roles only very recently. Remarkably, the most ancient ABEV repertoires were from dsRNA and minus-ssRNA viruses, suggesting they predated DNA viruses in evolution, a hypothesis we further test below.

Additional Support to the Early Origin of RNA Viruses

The proposal of viruses being a separate domain of life is not new (e.g. [292,293]). It has been the subject of intense debate in evolutionary biology (refer to [199,207,231,237,294-297] and references therein). Our analysis of protein domain structures suggests that there is a significant overlap in the proteomes of cells and viruses and also within the viral supergroup. This overlap identifies viruses as a unique 'fourth supergroup' along with Archaea, Bacteria, and Eukarya. However, formally placing viruses in the ToL is a daring task because many scientists even question the idea of viruses as living organisms mainly due to the lack of true viral metabolism and the inability to reproduce on their own [236,237]. However, the virocell concept [240,241] and the discovery of giant viruses [214-217] have furthered our understanding of the viral mode of life. Some have argued that the true living form of a virus is the intracellular virion factory that behaves like many other obligate intracellular parasites and is metabolically active [239]. Specifically, virocells produce viral gametes (virions) that are functionally analogous to cellular gametes of sexually reproducing species, which fuse during fertilization. These viral gametes can then fertilize (read infect) other cells (*sensu* [239]). In other words, virions are indeed metabolically inactive but are only a means to disseminate genetic information and complete the viral reproduction cycle. In turn, the virion factory or the transformed virocell represents the living form of viruses [239-241]. Thus, viruses should be considered 'living' organisms that simply survive via an atypical reproduction method that requires infecting a cell

(similar to obligate parasitism [242]). Moreover, the practice of modern phylogenetic analysis is to project genomic components of organisms onto the ToL and not their phenotypes; by that definition viral genomes are also part of the ToL [298]. In short, there is need to broaden our definitions of 'life' and abandon viewing virions as viruses (*sensu* [239]). Taken together, we argue that it is legitimate to build a universal ToL that includes viruses and truly describes the diversity of the living world.

To describe the evolutionary relationships between the proteomes of cells and viruses (taxa), we reconstructed conventional ToPs from the abundance and occurrence of 442 ABEV FSFs (phylogenetic characters). The ABEV Venn group included many FSFs of ancient origin (median *nd* ~0.4, Figure 4.5A) and the entire *abe* core (Figure 4.3A) and ancient FSFs in Ariadne's threads (Figure 4.4B). Importantly, these FSFs were particularly widespread in both cells and viruses, thus becoming the most appropriate FSF subset for reconstruction of rooted ToPs. Because biases in taxon sampling could influence tree reconstruction, we randomly sampled a set of 368 proteomes (taxa) from cells and viruses, including up to 5 viral species from each viral order or family and 34 proteomes corresponding to only 'free-living' organisms in Archaea, Bacteria, and Eukarya. Recently sequenced proteomes of giant viruses [214-217] and their virophages [234,299,300] were all part of the sampled viral sub-group. This yielded a total dataset of 368 proteomes, including 266 viruses (92 dsDNA, 15 ssDNA, 42 dsRNA, 90 plus-ssRNA, 12 minus-ssRNA, 5 ssRNA-RT, and 10 dsDNA-RT) and 102 cellular organisms.

ToLs describing the evolution of proteomes (ToPs). The rooted phylogeny dissected proteomes into four groups (Figure 4.6A). Remarkably, viruses formed a distinct paraphyletic group at the base of the ToP that was distinguishable from cells by 76% BS. In turn, archaeal organisms were clustered paraphyletically in the more basal branches (black circles), while Bacteria and Eukarya formed strong monophyletic groups (blue and green circles) supported by 66% and 100% BS values, respectively. (Figure 4.6A). This topology supported an ancient origin of both viruses and Archaea and a sister relationship between Bacteria and Eukarya, which goes against some gene sequence-based phylogenies [155,165,180] but is congruent with a number of structure and function-based studies (discussed elsewhere [20,27,83,84,154,301]).

Within the viral supergroup (or the 'fourth' domain), the most basal taxa corresponded to RNA and retrotranscribing viruses. These included well-known dsRNA viral families that

possess segmented genomes such as *Birnaviridae*, *Partitiviridae*, and *Picobirnaviridae* (2 segments), *Chrysoviridae* and *Quadriviridae* (4 segments), and *Reoviridae* (10-12 segments). Interestingly, *Nodaviridae* that possess bipartite genomes (i.e. 2 segments) and 'capsid-less' *Narnaviridae* (both plus-ssRNA) also occupied the most basal positions in the ToP along with dsRNA and dsDNA-RT viruses. Other very ancient viral groups included retrotranscribing (*Caulimoviridae*, *Hepadnaviridae*, and *Retroviridae*), ssDNA (*Anelloviridae* and *Inoviridae*), dsDNA (*Plasmaviridae* and *Polydnaviridae*), ambisense arenaviruses, and minus-sense influenza viruses (Figure 4.6A). It has been hypothesized that retrotranscribing viruses likely mediated the transition from an ancient RNA to the modern DNA world [203]. Remarkably, retrotranscribing viruses originated prior to the DNA viruses in our tree, thus validating the hypothesis (Figure 4.6A). Another interesting position was of polydnaviruses that exist as 'symbionts' of endoparasitic wasps [302]. Interestingly, these viruses also encode segmented dsDNA genomes! These observations suggest an ancient presence of segmented viral genomes (mostly RNA) and the late appearance of 'capsid-encoding' and DNA viruses.

The ToP also recovered some other well-known relationships. For example, *Flavivirus* (*Flaviviridae*) and *Alphavirus* (*Togaviridae*) genera were grouped together suggesting their close evolutionary association (66% BS). In fact, alphaviruses were initially classified under *Flaviviridae* by ICTV but were later assigned their own genera within *Togaviridae*. Both viral families show striking similarities in virion architecture (enveloped and spherical) and genome replication strategies (monopartite linear plus-ssRNA). Similarly, *Polyomaviridae*, *Closteroviridae*, *Coronaviridae*, and many others also formed individual monophyletic groups. Another largely unified group was of filamentous dsDNA archaeoviruses; *Rudiviridae* and *Lipothrixiviridae* that have been classified under order '*Ligamenvirales'* [303]. Similarly, viral families within the Nucleocytoplasmic Large DNA viruses group (*Poxviridae*, *Phycodnaviridae*, *Ascoviridae*, *Asfarviridae*, *Iridoviridae*, and *Mimiviridae*) formed a paraphyletic group at the very derived positions. This group also included the recently discovered pandoraviruses and pithoviruses and the oddly placed single bacteriophage (*Myoviridae*). The close grouping of all giant viruses supports the proposal of a novel viral order '*Megavirales'* [304] and a previous reconstruction [14].

However, *Herpesviridae* and *Caudovirales* that share the HK97 capsid protein fold did not form a single group [227], but they were in close proximity (Figure 4.6A). In turn,

Adenoviridae and *Tectiviridae* that belong to the PRD1/Adenovirus lineage were clustered closely. Similarly, *Totiviridae* and some *Reoviridae* of the BTV-like lineage occupied basal positions. Some members of the Picornavirus-like lineage (e.g. *Luteoviridae*, *Caliciviridae*, *Picornaviridae*) and retrotranscribing viruses also clustered together but clear-cut structure-based viral lineages did not materialize in the ToP. We emphasize that our phylogenies consider both the structural (capsid) and non-structural (polymerases and others) proteins in characterizing proteomes and give a global proteomic view as opposed to focusing on a single phylogenetic character (i.e. capsid). Other discrepancies also existed with regard to viral families defined by ICTV that did not form unified groups. However, ICTV classifications are subject to revisions and do not always yield evolutionarily informative classifications. In light of these, the ToP reconstructed from the abundance of conserved FSF domains presents a 'third' and global view of the evolutionary relationships of viruses, which adds deep lineage relationships to the structure-based and ICTV classifications.

Interestingly, most basal branches were populated by spherical or filamentous virions (two of the simplest designs from a tensegrity point-of-view). They gradually become more decorated with additional features such as spikes and glycoproteins (retroviruses) in spherical virions and rod-like designs (inoviruses) likely evolving from filamentous versions (Figure 4.6A). Perhaps, the rods and spheres combine to form head-tail morphotype so abundant in prokaryotic viruses. Thus, mapping of virion morphotypes onto the ToP likely hints towards the origin of viruses from a limited number of structural designs. However, we caution that morphological similarities may also stem from convergent evolution. At this point, we lack evidence to confirm homologies between different virion morphotypes. Nevertheless, the early appearance of spherical and filamentous virions harboring segmented RNA genomes is remarkable and worthy of further attention. Finally, the global phylogenetic relationships of cellular organisms that were used as reference supported previous reconstructions of this kind, including an early paraphyletic origin of Archaea and a sister relationship between Bacteria and Eukarya [43,301].

While the evolutionary groupings of viruses in the ToP may be subject to phylogenetic artifacts, most likely taxon sampling, an early origin of segmented RNA and capsid-less viruses is noteworthy and was also predicted earlier from the comparative genomics and ToD analysis. Remarkably, a tree of viruses (ToV) reconstructed from the *abe* core FSFs (Figure D4) further

confirmed an early origin in RNA viruses. While patterns of distribution of replicon types were not entirely clear-cut, there was clear enrichment of RNA viral proteomes at the base of the ToP, specifically minus-ssRNA and dsRNA viruses. This tree was poorly resolved partly due to the limited number of phylogenetic characters that were used to distinguish proteomes and largely due to the patchy distribution of *abe* FSFs in viral proteomes (a consequence of reductive evolution in viruses). Finally, grouping viruses by host type (i.e. archaeoviruses, bacterioviruses, and eukaryoviruses) did not yield three independent groups suggesting that viruses, regardless of the host type, could be structurally (and evolutionarily) more related to each other (Figure D5). It also suggests that viruses can jump hosts (e.g. SARS and Ebola viruses, loss of RNA viruses in prokaryotes [212]) and thus inferring evolutionary relationships based on virus-host preferences may be misleading (Chapter 6).

Phylogenomic networks describing the evolution of proteomes. Typically, viral proteomes encode far less proteins and in lower abundance relative to the proteomes of cellular organisms (except for some giant viruses). To account for such differences and to test if the phylogeny in Figure 4.6A was not influenced either by HGT or our choice of phylogenetic model, we also employed FSF occurrence into distance-based phylogenomic networks reconstruction (Figure 4.6B). Distance-based methods employ far simpler models to calculate a taxonomy for given taxa and can be used to test conflicts between phylogenetic model and output tree. The resulting topology still favored a 'tree-like' structure (Figure 4.6B) suggesting that the phylogeny of Figure 4.6A was not influenced by processes that could artificially increase genomic abundance. Moreover, none of the viral proteomes clustered with their hosts (e.g. plant RNA viruses did not group with plants) indicating that the predicted cellular nature of viruses was not due to HGT from their hosts but was likely a result of ancient co-existence. Importantly, the phylogenomic network retained the majority of evolutionary relationships defined earlier by the ToP but also recovered a closer grouping of herpesviruses with *Podoviridae* (*Caudovirales*) that was not so clear in the ToP derived from genomic abundance, supporting the proposal that the two viral groups are closely-related [227,277].

ToLs derived directly from the age of protein domains. We also used an innovative approach of multidimensional scaling to study the evolution of cells and viruses, the evoPCO (Figure 4.7A). The evoPCO method combines the power of cladistic and phenetic approaches by calculating principal coordinates directly from temporal evolutionary distances between the

proteomes of species (see Methods). The distance between proteomes reflects phylogenetic dissimilarity in the age of the FSF domain repertoires (i.e. *nd* values) and can be displayed in 3D temporal space, assuming that the age of an FSF is the age of the first instance of that FSF appearing in evolution. Because, proteomes are biological systems that are made up of component parts (i.e. FSFs in this case) but describe cellular organisms and viruses, each component (regardless of its abundance) contributes an age to the overall age of the cellular or viral system. This factor, when taken into account, resulted in a powerful projection of a multidimensional space of proteomes onto a 3D temporal space that allows visualization of evolutionary relationships.

Remarkably, the evoPCO revealed four clear clouds of proteomes in temporal space that corresponded to viruses and the three cellular superkingdoms (Figure 4.7A). The first three coordinates explained ~85% of the total variability. Using the previously reconstructed proteome of LUCELLA as a reference point [38], we inferred viruses as the most ancient supergroup, followed by Archaea, Bacteria, and Eukarya, in that order (Figure 4.7A). This topology supports earlier results from the comparative genomic and phylogenomic analyses, adding a third line of evidence in support of the early origin of viruses. Remarkably, *Lassa virus* that belongs to *Arenaviridae* and harbors segmented RNA genomes appeared at the most basal position of the evoPCO plot, supporting the early origin of segmented RNA viruses recovered earlier in ToPs (Figure 4.6A). Some giant viruses appeared more derived supporting their ancient co-existence with cells [14,195]. The topology and ordering of proteomes in evoPCO was further supported by a distance-based NJ tree (Figure 4.7B) reconstructed directly from the temporal distance matrix, which retained the cohesive and ancient nature of the viral supergroup. The NJ tree made explicit the early origins of RNA viral families and was largely congruent with the ToP recovered earlier (Figure 4.6A), validating the power of the evoPCO strategy.

Aridane's threads traced in evolutionary time. Our Ariadne's threads (Figs. 4.4B and 4.8A) further dissected viral origins, again pointing toward an early origin of RNA viral replicons. We traced FSF domain ages onto the threads of FSFs shared between viral subgroups (Figure 4.8A). The oldest domains were spread in a transect that unified minus-ssRNA, plus-ssRNA and dsRNA proteomes. This pattern was clearly evident in violin plots that describe FSF age in the threads along the early timeline of domain evolution (*nd* < 0.3) (Figure 4.8B). Once

again, the proteomes of minus-ssRNA viruses were particularly enriched in ancient domains, suggesting that perhaps single stranded RNA was involved in virocell origins (read below).

Arguments in Favor of RNA-based Proto-virocells

The early appearance of RNA viral replicons recovered consistently from the comparative and phylogenomic approaches of this study is a significant finding that supports the general belief that RNA came before DNA. The findings also support the existence of RNA genomes in LUCA, and are in line with previous phylogenomic reconstructions [38]. The ubiquity of use of RNA primers in DNA synthesis and synthesis of deoxyribonucleotide precursors of DNA from the ribonucleotide precursors of RNA [305] is additional support for this argument. Remarkably, the very basal viral groups in ToPs and ToVs (Figs. 4.6A and D4) included minus-ssRNA influenza viruses and families of dsRNA viruses that harbor segmented genomes. Influenza virus genomes typically contain 6-8 RNA segments and evolve by random genetic drift or by the reassortment of genome segments with other co-infecting influenza viruses. Thus, it is likely that the earliest cellular forms, including proto-virocells, possessed segmented RNA genomes that often 'mated' by combining with other RNA segments (for example see 'multiplicity reactivation' [306]). This scenario is compatible with Woese's proposal that the earliest cells stored genes in the form of segmented RNAs [307]. The principle of continuity dictates that a possible shift from RNA to DNA was gradual and was likely mediated by retrotranscribing viruses (e.g. see Figure 4.6A).

Another argument in favor of the early origin of minus-ssRNA viruses is the genomic makeup and replication strategy of ambisense viruses (arenaviruses) that appeared prior to influenza viruses in the reconstructed universal ToLs (Figure 4.6A). These viruses possess two genomic RNA segments (L and S). One region of each segment is transcribed from 5'-3' (i.e. positive polarity) while the other halve is read from 3'-5' (negative polarity). Importantly, the genes on the negative-sense halve are transcribed first and regulate the emergence of mRNA transcripts. The obvious evolutionary advantage is the synthesis of multiple mRNA transcripts from a single negative-strand. In turn, positive-sense RNA viruses only harbor a single mRNA and produce polyproteins that are later cleaved into individual products. Thus, the ability to read negative-sense RNA into many mRNAs must be a significant evolutionary breakthrough as it would be the first step towards efficient translation. Negative-sense and dsRNA viruses that

possess segmented genomes and utilize these transcription and replication strategies must therefore be at the forefront of cellular evolution and likely led the transition to full-length plus-sense RNA and later DNA genomes.

Interestingly, many RNA virus genomes encode specialized structures to ensure compatibility with cellular machinery. For example, polioviruses encode an internal ribosome entry site (IRES) at the 5'-UTR to initiate translation from the middle of mRNA (i.e. cap-independent protein translation). The *Taura syndrome virus* encodes a 'tRNA-mRNA like' structural element that binds to the aminoacyl site of the ribosome to initiate protein translation [308]. It was experimentally shown that the IRES sites of several viruses are structurally similar to 'tRNA-like' domains [309]. tRNA-like structures have also been discovered in a large number of other viruses such as *Bacteriophage Qβ* and retroviruses where they serve as templates and primers for replication, respectively. A similar role is likely achieved in ambisense arenaviruses by the conserved 19 bp nucleotide sequences at each terminus of the genomic segments that facilitates in replication and transcription and forms a hairpin secondary structure [310]. Interestingly, it was proposed that tRNA evolved from the duplication of hairpin coding sequence [311]. The presence of these tRNA-like structures in RNA viruses is compatible with the 'genomic tag' hypothesis presented more than 25 years ago [312,313]. Under this hypothesis, tRNA-like structures were encoded by RNA genomes of a primordial 'RNA world' (read RNA genomes of proto-virocells) that served as replication initiation sites and tagged genomic RNAs for either replication or other roles. The widespread presence of such structures in extant RNA viruses supports the idea that RNA replicons (more likely the segmented versions) should be the progenitors of the 'DNA world' and modern protein synthesis. However, we stress that contemporary RNA viruses may not be the direct descendants of the earliest RNA replicons. Instead, they simply reflect the earliest replication strategies that were 'tried' by the first virocells. The remarkable diversity of viral replication strategies (i.e. the seven Baltimore groups) is strong support for this experimentation.

Significance

The origin of viruses has been mysterious, especially because of their diverse and patchy molecular and functional makeup. While numerous hypotheses, mainly theoretical, explain viral origins, none is backed by substantive data. In this study we explored the proteomic makeup of

thousands of proteomes sampled from cells and viruses and studied their evolution taking advantage of the wealth of protein structural and functional data that is available. Remarkably, we established an ancient origin of the viral supergroup despite the extremely reduced nature of their proteomes and the possible existence of widespread episodes of horizontal transfer of genetic information. We found that viruses fuel cellular diversity while being themselves the subject of strong lifestyle-driven reductive evolutionary pressures.

Our results suggest that viruses always interacted with cells but did so in a different manner in early evolution. Viruses seem to have originated from proto-virocells that co-existed with proto-cells ancestors of LUCELLA. While infection of modern day virocells often results in virion synthesis and cell lysis [240,241], the proto-virocell genomes co-existed with the intracellular environment and reproduced without lysis. The evidence for such co-existence comes from the widespread sharing of a large number of ancient proteins between cells and viruses infecting hosts separated by large evolutionary distances. It also explains the origin of a large number of unique proteins in viruses that lack cellular homologs, as they would originate continuously in a cellular environment throughout evolutionary timeline (i.e. in both proto-virocells and modern virocells). There could be many other proto-cells residing in that ancestral community but eventually became extinct. The only known survivors are the three superkingdoms and their obligate parasites, the viral supergroup. We argue that segregation into hosts and parasites is a natural outcome of any competitive system.

Proteomes of all seven kinds of viral replicons were enriched in ancient FSFs (Figure 4.5C). Given the massive replicon type diversity seen in modern viruses, it is likely that all kinds of replication strategies were utilized in proto-virocells. A logical outcome of this experimentation would be the discovery of many key replication-associated proteins and perhaps DNA itself in the virus world (an idea previously put forward by Forterre [194,203,314]). The shift to the ‘viral’ mode of life occurred once reductive evolutionary modes and specialization forced proto-virocells to forfeit most of the translation machinery that was unfolding. The late appearance of the ribosomal biosynthetic apparatus in evolution is compatible with recent phylogenomic reconstruction analyses that revealed coevolution of ribosomal proteins and RNA [97]. The prolonged pressure of genome and vesicle-like reduction fundamentally transformed the two main components of proto-virocells (compartments and viral replicons), first by intracellularly segregating replicons in viral factory-like compartments (likely ancestors of the

eukaryotic nucleus) and then by creating virion particle structures that would efficiently propagate genetics. The rather late appearance of capsids thus marks the unfolding of viral parasitism, extracellular lifecycles, and modern virocells. Under this scenario, plasmids and other selfish genetic elements also originated from proto-virocells but did not acquire capsids and remained tightly integrated with the emerging ribocellular make up. This biphasic cell-like and parasitic model of viral evolution should change the way we view viruses, from parasitic nuisances to helpful entities of cellular innovation.

Reconciliation with Existing Hypotheses

Evolution is a gradual process that takes billions of years to unfold. It is bidirectional and can move both from simple to complex (multicellular organisms) and complex to simple (many obligate intracellular parasites, organelles, and viruses). We emphasize that the traditional views regarding viral evolution can be reconciled and improved in light of data that we have generated. The crucial helping factor would be to realize that modern cells and modern viruses are simply evolved forms of ancient cells and viruses over billions of years (*sensu* [194]). This distinction helps to understand the complex interaction between cells and their apparent nemesis (viruses) and enables us to present a viral origin hypothesis that is parsimonious with genomic data.

We first refute the idea of a pre-cellular origin of viruses (see [207] for a new version). The virus-first hypothesis suggests an early origin of self-replicating viral replicons predating the origin of cells. It is mainly supported by the finding that some key viral proteins lack cellular homologs but are shared by many RNA and DNA viruses (e.g. the ‘jelly-roll’ capsid). However, viruses are continuously creating new genes during virocell lifecycles and thus many unique viral proteins maybe of very recent origin (e.g. V group in Figure 4.5A). Moreover, structural analogs of ‘jelly-roll’ folds and others have been detected in cellular proteomes (e.g. histone chaperones and encapsulins). Also, viruses are tightly associated with proteins (capsids) in their make up and must replicate in an intracellular environment thus necessitating the need of a cell before virus. Fossil evidence also supports that cells originated very early in evolution [315,316]. Thus, the virus-first hypothesis lacks explanatory power and is not compatible with virus biology. The escape hypothesis, on the other hand, associates the origin of viral genomes with modern cellular genomes. However, the massive number of viral proteins that lack cellular homologs and numerous instances of virus-to-cell gene transfer challenge the hypothesis.

Viruses seem to have their own unique identity and the very large number of unique viral proteins is more parsimonious with archaic proto-virocells co-existing with the cellular ancestors. This leaves us with the reduction hypothesis that suggests that viruses reduced from cells. We argue that this hypothesis is best compatible with genomic data, especially when one invokes reduction from ancient cells and not from modern cells. It is logical as nearly all known obligate symbionts and parasites in the three superkingdoms of life follow a similar route [268] and is supported by the discovery of giant viruses [214-217] that overlap parasitic cells in both genome and particle size. The reductive scenario explains the origin of unique viral proteins that lack cellular homologs simply by invoking an additional sibling of the cellular ancestors, the proto-virocell. Remarkably, prokaryotic protein compartments (carboxysomes and encapsulin protein shells) may possess protein folds that were once utilized by ancient viruses to infect ancient cells.

Some Technical Considerations

We focused on the abundance and occurrence of FSF domains in proteomes. It can be argued that abundance of some folds could be artificially increased by non-vertical evolution such as HGT or decreased due to incomplete or biased sampling or simply due to evolutionary bottlenecks (e.g. loss of an ancient fold from the ancestor of a superkingdom). However, we note that the abundance-based approach is relatively more robust against non-vertical evolutionary forces, mainly HGT. The effect of HGT-related artificial increases in genomic abundance for ancient FSFs would be almost negligible (as those already have high abundance count in genomes). In turn, HGT gain of some of the recently evolved FSFs that are present in genomes with low count (e.g. 1-2 per genome) could be significant but would only affect the very derived branches of the ToD. Moreover, occurrence-based and abundance-based analysis provided largely congruent results, suggesting that both parameters of the structural census carry similar signatures of the evolutionary process.

In terms of character polarization, it could be argued that viruses with very small proteomes can be artificially attracted to basal branches of ToPs making the construction of a universal ToL problematic. This interpretation however is erroneous since polarization also involves spread in the nested lineages of the ToL and is only applied *a posteriori*, allowing gains and losses throughout branches of the tree [317]. We note that assumptions of character

polarization comply with Weston's generality criterion of phylogenetic rooting [45,46] and are consistent with the proposal of a simpler progenote organism (community) at the beginning of evolution. A number of theoretical arguments and experimental evidence support the assumption that ancient genes have more time to accumulate and spread in diversifying lineages. For example, the 'P-loop containing NTP hydrolases' FSF (c.37.1) includes ubiquitous and highly abundant proteins that are involved in membrane transport and metabolic processes. There is general agreement that these proteins evolved first in evolution. FSF c.37.1 was also the first to appear in our ToD and this result was consistently recovered in many previous phylogenomic reconstructions (e.g. [5]). We also note that the ancestry of FSFs in ToDs depends upon the 'profile' distribution of FSFs in proteomes. For example, some immunoglobulin superfamily domains are very abundant in some eukaryotes. Despite their very high abundance in some organisms, they are not the most ancient FSFs in our ToDs. This implies that both abundance and spread of FSFs determine the position of FSFs in timelines derived from phylogenetic trees. Still, comparing phylogenies obtained from occurrence and abundance counts of FSFs can experimentally validate polarization (e.g. [317]). For example, distance-based phylogenies yield topologies that are congruent with ToPs (Figs. 4.6 and 4.7), and similar conclusions were strongly supported by comparative genomic experiments. Thus, the ancient history of the viral supergroup should be considered reliable unless strong evidence is presented to suggest otherwise.

We also developed evoPCO to support other tree reconstructions, a novel approach that combines the cladistic power of tree reconstruction from shared and derived characters with statistical and phenetic approaches of ordination. The evoPCO recovered robust phylogenetic relationships between cells and viruses, revealing the primordial origin of viruses and the very early appearance of Archaea. These relationships validated the ToL polarization scheme used in ToP reconstruction by showing that FSFs ages from ToDs in evoPCO provided a same 'evolutionary arrow', also matching comparative genomic predictions. Thus, several independent lines of evidence mutually support character polarization and evolutionary statements for building rooted ToLs.

It is particularly noteworthy that the new evoPCO methodology is free from typical artifacts that complicate ToL reconstruction, most importantly character independence [37,95]. ToLs are generally built from nucleotide or amino acid site information in nucleic acid or protein

sequences, which are generally not independent from each other because of the mere existence of molecular structure [271]. This violates the phylogenetic requirement of character independence, unless suitable representations of structure-based dependencies are incorporated into the evolutionary tree-building model. In contrast, FSF ages used in evoPCO were calculated from a ToD, a tree derived from domain abundance counts in proteomes, which are used as characters. Since proteomes generally evolve independently from each other (except for symbiotic or trophic interactions) and any possible interaction between them occurs at levels of organization that are much higher than molecular structure, ToDs (and temporal information they provide) are therefore impervious to the need to budget molecular structural dependencies of characters in evolutionary models.

Limitations of Our Findings

Our conclusions rely on the accuracy of HMMs to detect FSF domain structures in protein sequences and the current power of SCOP, an influential gold standard of protein classification. The structural census could be the subject of biases in the genomes that have been sequenced so far and our ability to appropriately survey viral and cellular biodiversity. While phylogenomic reconstructions depend upon the choice of phylogenetic model and search strategy for optimal trees, our experience with these methodologies has shown that in general phylogenetic reconstructions are reliable. We therefore assume retrodiction statements are not biased by preconceptions of modernity in the extant features that are studied. We also note that our focus is on protein domain structure and not protein sequence. We therefore avoid the time-erasing effect of mutations and the confounding convergent effects of historical patchworks present in multidomain protein sequences, which represent a substantial fraction of every proteome that has been sequenced [37]. Thus, our analysis provides an evolutionarily deep 'structural' view that as expected is not always in line with the shallow 'sequence' view of viral evolution. This fact should be taken into consideration when interpreting our conclusions. Finally, we stress that our conclusions are the 'most likely' scenarios inferred from both comparative genomics (e.g. Venn diagrams and *f*-values) and phylogenomic approaches (ToDs, ToPs and evoPCO). Studying viral and cellular evolution is a difficult problem complicated by many logical and technical considerations. In light of these, we hope that our study will initiate further discussion on this topic and that a consensus regarding viral evolution will be reached in the near future benefitting both viral biology and taxonomy.

Conclusions

Viruses display extraordinary diversity in their molecular makeup, morphologies, and replication strategies. Despite remarkable levels of variability, viral protein structures retain traces of their evolutionary history, which can be recovered using advanced bioinformatics approaches. Applying both comparative genomics and cladistics approaches, we uncovered remarkable trends in the evolution of the viral supergroup. A large number of FSF domains were detected in the viral proteomes including those that lacked cellular homologs. Viruses and cells shared numerous FSFs, most of which were of ancient origin. Viruses with different replicon types and infecting distantly related hosts also shared many structural and informational FSFs. FSFs shared with viruses were significantly more widespread in cellular proteomes suggesting viruses mediate cellular diversification. Remarkably, structural relatives for some capsid-associated FSFs were detected in cells but were not widespread. These FSFs uniquely link the viral and cellular worlds. Structural phylogenomic analyses confirmed the early appearance of the viral supergroup and identified viral RNA as the primary genetic material of earliest protocells. Global phylogenomic reconstructions of FSF and proteome history showed that the make up of the viral supergroup was very ancient and was reductively shaped by gene loss, which started very early in evolution and finally resulted in viral adaptation to obligate parasitism. Observations strongly support an early common history and coexistence of viruses and ancestral cells and the evolutionary cohesiveness of the viral supergroup.

Figures

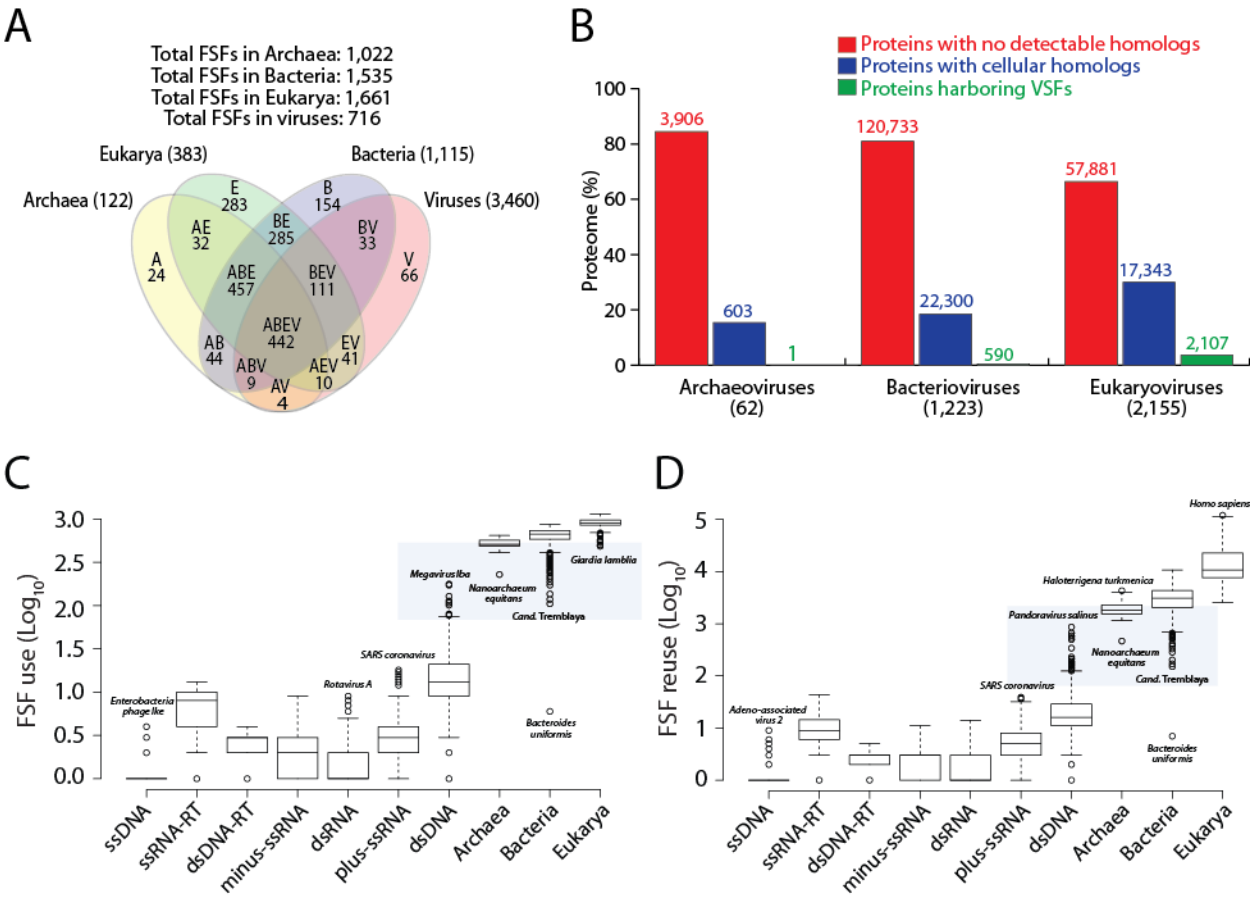

Figure 4.1 FSF sharing patterns and make up of cellular and viral proteomes. A) The Venn diagram defines 15 mutually exclusive groups based on the distribution of 1,995 FSFs in 5,080 proteomes sampled from cells and viruses. Numbers in parentheses indicate total number of proteomes that were sampled from Archaea, Bacteria, Eukarya, and viruses. **B)** Barplots comparing the proteomic composition of viruses infecting the three superkingdoms. Numbers in parenthesis indicate total number of viral proteomes in each group. Numbers above bars indicate total number of proteins in each of the three classes of proteins. VSFs listed in **Table 4.1**. **C, D)** FSF use (i.e. number of unique FSFs in a proteome) and reuse (total number of FSFs in a proteome) for proteomes in each viral subgroup and the three superkingdoms. Use and reuse are given in logarithmic scale. Important outliers are labeled. Shaded regions highlight the overlap between parasitic cells and giant viruses.

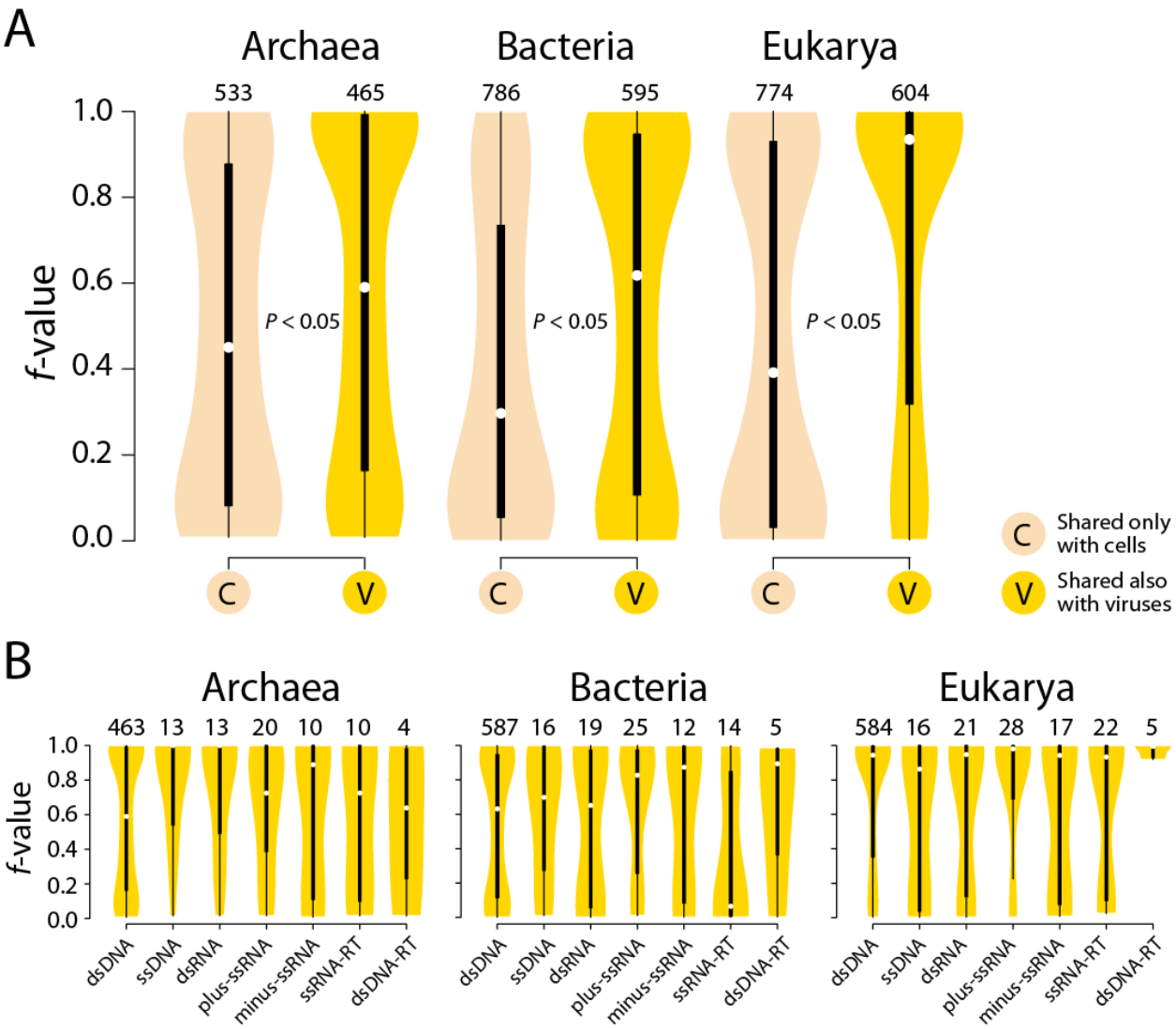

Figure 4.2 Spread of viral FSFs in cellular proteomes. A) Violin plots comparing the spread (*f*-value) of FSFs shared and not shared with viruses in archaeal, bacterial, and eukaryal proteomes. **B)** Violin plots comparing the spread (*f*-value) of FSFs shared with each viral subgroup in archaeal, bacterial, and eukaryal proteomes. Numbers on top indicate total number of FSFs involved in each comparison. White circles in each boxplot represent group medians. Density trace is plotted symmetrically around the boxplots.

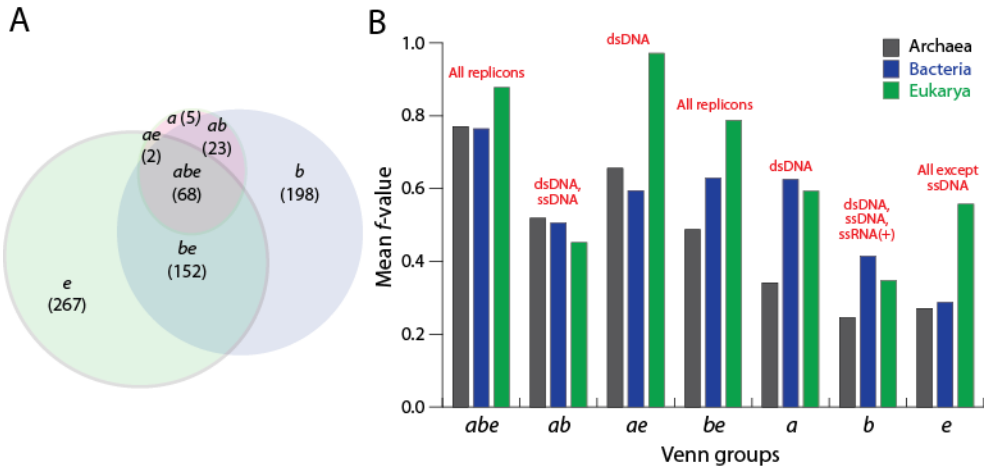

Figure 4.3 Virus-host preferences and FSF distribution in viruses infecting different host organisms. Virus host information was taken from NCBI Viral Genomes Resource [254]. Hosts were classified into Archaea, Bacteria, Protista (animal-like protists), Fungi, Plants (all plants, blue-green algae, and diatoms), Invertebrates and Plants (IP), and Metazoa (vertebrates, invertebrates, and human). Host information was available for 3,440 out of 3,660 viruses that were sampled in this study. Two additional ssDNA archaeoviruses were added from literature [318,319]. **A)** Venn diagram shows the distribution of 715 (out of total 716) FSFs that were detected in archaeo-bacterio- and eukaryoviruses. Host information for *Circovirus like genome RW_B* virus encoding 'Satellite viruses' FSF (b.121.7) was not available. **B)** Mean *f*-values for FSFs corresponding to each of the seven Venn groups defined in **A** in archaeal, bacterial, and eukaryal proteomes. Values were averaged for all FSFs in each of the seven Venn groups. Text above bars indicates how many different viral subgroups encoded those FSFs.

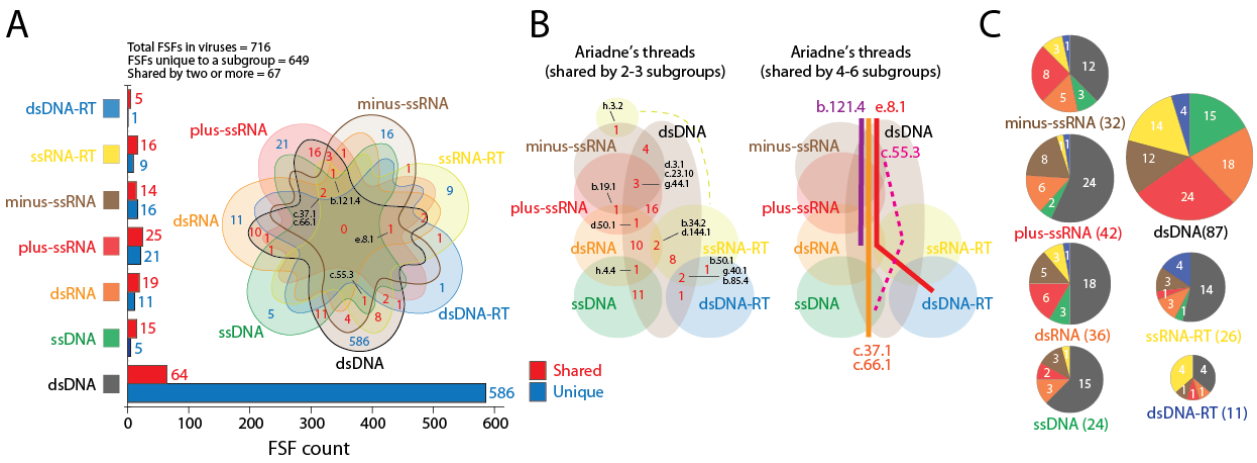

Figure 4.4 Distribution of FSFs within the viral supergroup. A) Total number of FSFs that were either shared or were uniquely present in each viral subgroup. A 7-set Venn diagram makes explicit the 127 (2^7-1) combinations that are possible with seven groups. **B)** Ariadne's threads give the most parsimonious solution to encase all highly shared FSFs between different viral subgroups. Threads were inferred directly from the 7-set Venn diagram. FSFs identified by SCOP *css*. **C)** Number of FSFs shared in each viral subgroup with every other subgroup. Pie charts are proportional to the size of FSF repertoire in each viral subgroup.

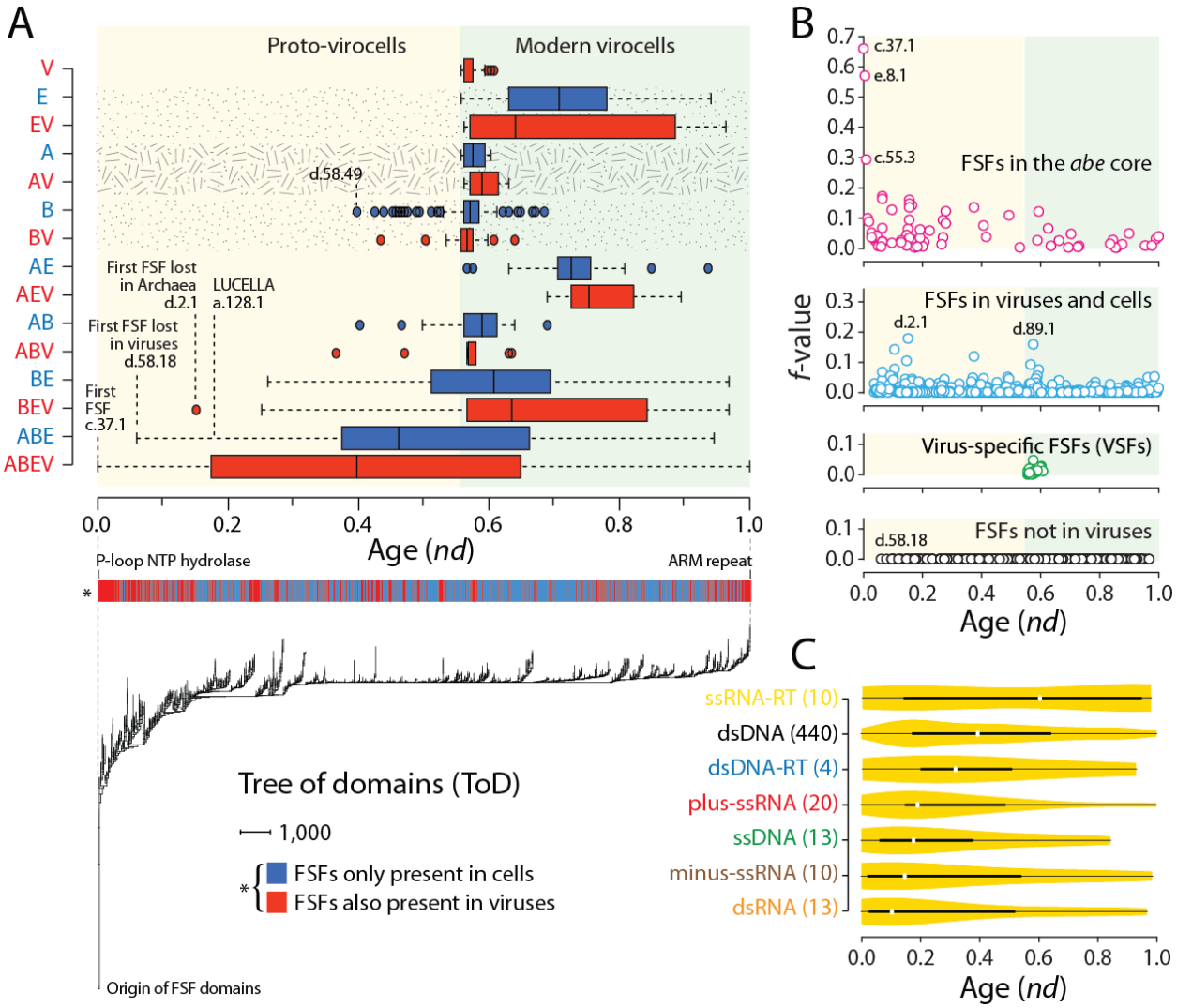

Figure 4.5 Phylogenomic analyses of FSF domains. A) The ToD describes the evolution of 1,995 FSF domains (taxa) in 5,080 proteomes (characters) (tree length = 1,882,554; Retention Index = 0.74; g_1 = -0.18). The bar on top of the ToD is a simple representation of how FSFs appeared in evolutionary time (i.e. *nd*). FSFs were labeled blue for cell-only and red for those either shared with or unique to viruses. The boxplots identify the most ancient and derived Venn groups defined in Figure 4.1A. Two major phases in the evolution of viruses are indicated in different background colors. Patterned area highlights the appearances of AV, BV, and EV soon after A, B, and E, respectively. FSFs identified by SCOP *css* (see text for description). **B)** Viral FSFs plotted against their spread in viral proteomes (*f*-value) and evolutionary time (*nd*). FSFs identified by SCOP *css* (see text for description). **C)** The distribution of ABEV FSFs in each viral subgroup along evolutionary time (*nd*). Numbers in parenthesis indicate total number of ABEV FSFs in each viral subgroup. White circles indicate group medians. Density trace is plotted symmetrically around the boxplots.

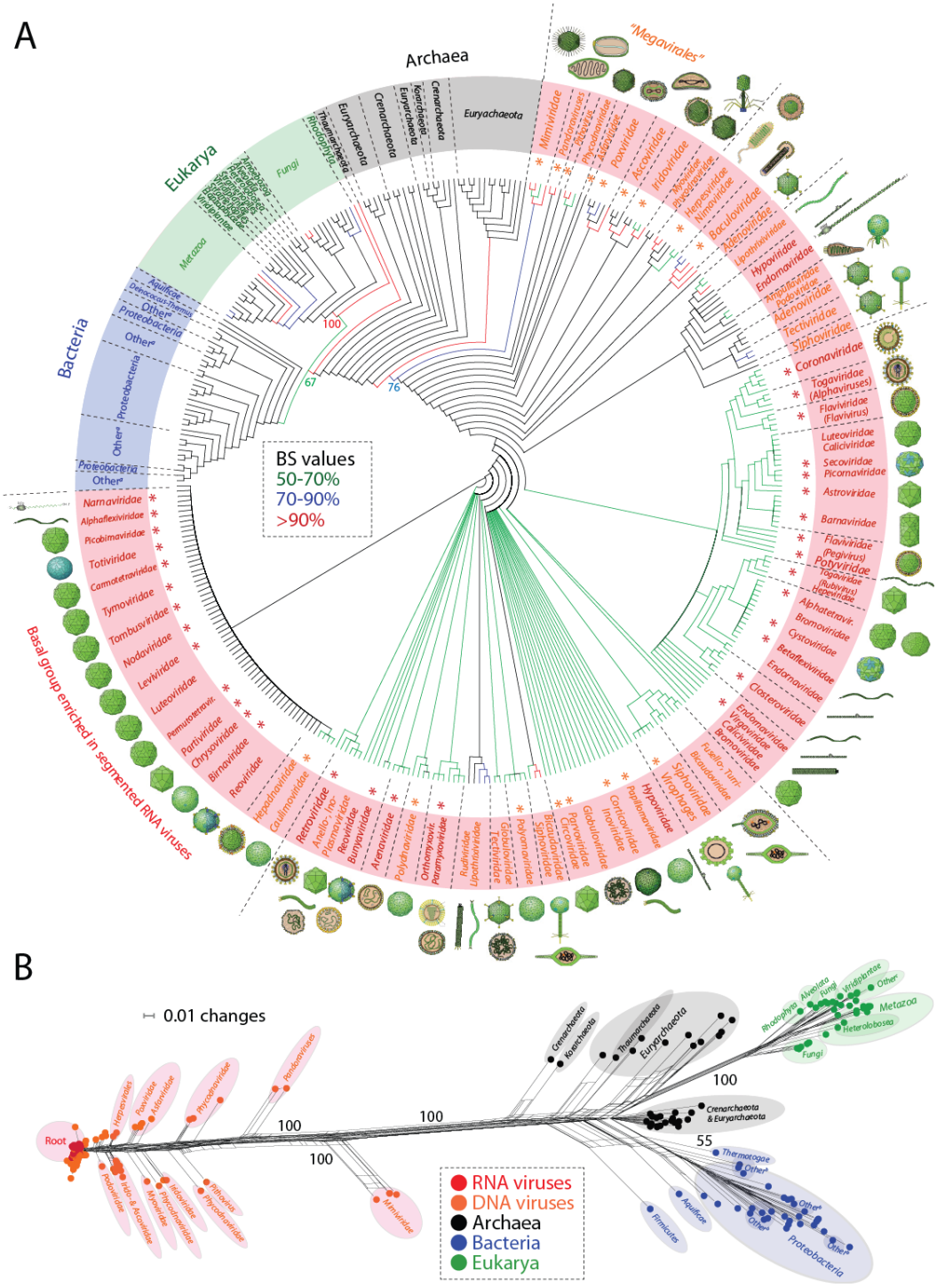

Figure 4.6 Evolutionary relationships between cells and viruses. A) A ToP describes the evolution of 368 proteomes (taxa) that were randomly sampled from cells and viruses and were distinguished by the abundance of 442 ABEV FSFs (characters) (Tree length = 45,935; Retention Index = 0.83; g_1 = -0.31). All characters were parsimony informative. Differently colored branches represent BS support values. Major groups are identified. Viral genera names are given inside parenthesis. The viral order "*Megavirales*" is pending approval by ICTV and hence written inside quotes. Viral families that form largely unified or monophyletic groups are identified by asterisk. Virion morphotypes were mapped to the ToP. Pictures taken from ViralZone web resource [320]. No picture was available for *Turriviridae*. [a]Actinobacteria, Bacteroidetes/Chlorobi, Chloroflexi, Cyanobacteria, Fibrobacter,

Firmicutes, Planctomycetes, and Thermotogae. **B)** A distance-based phylogenomic network reconstructed from the occurrence of 442 ABEV FSFs in randomly sampled 368 proteomes (Uncorrected P distance; Equal-Angle; Least-square fit = 99.46). Numbers on branches indicate BS support values. Taxa were colored for easy visualization. Important groups are labeled. [b]Actinobacteria, Bacteroidetes/Chlorobi, Chloroflexi, Cyanobacteria, Deinococcus-Thermus, Fibrobacter, Firmicutes, and Planctomycetes.

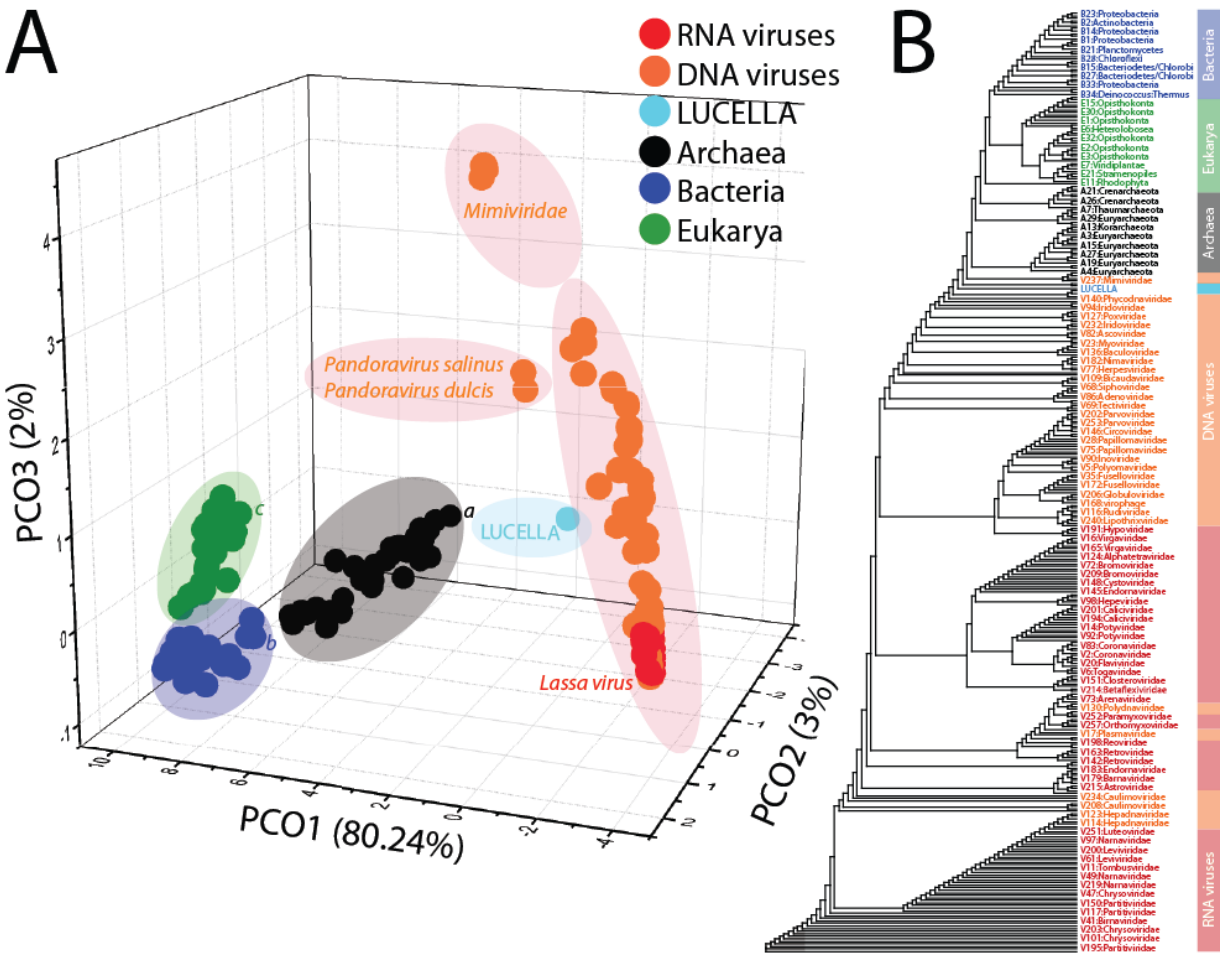

Figure 4.7 Evolutionary history of proteomes inferred from numerical analysis. A) Plot of the first three axes of evoPCO portrays evolutionary distances between the cellular and viral proteomes. The percentage variability explained by each coordinate is given in parenthesis on each axis. The proteome of LUCELLA (retrieved from [38]) was added as an additional sample to infer the direction of evolutionary splits. [a]*Ignicoccus hospitalis*, [b]*Lactobacillus delbrueckii*, [c]*Caenorhabditis elegans*. **B)** A distance-based NJ tree reconstructed from the occurrence of 442 ABEV FSFs in randomly sampled 368 proteomes. Each taxon was given a unique tree Id. Taxa colored for easy visualization.

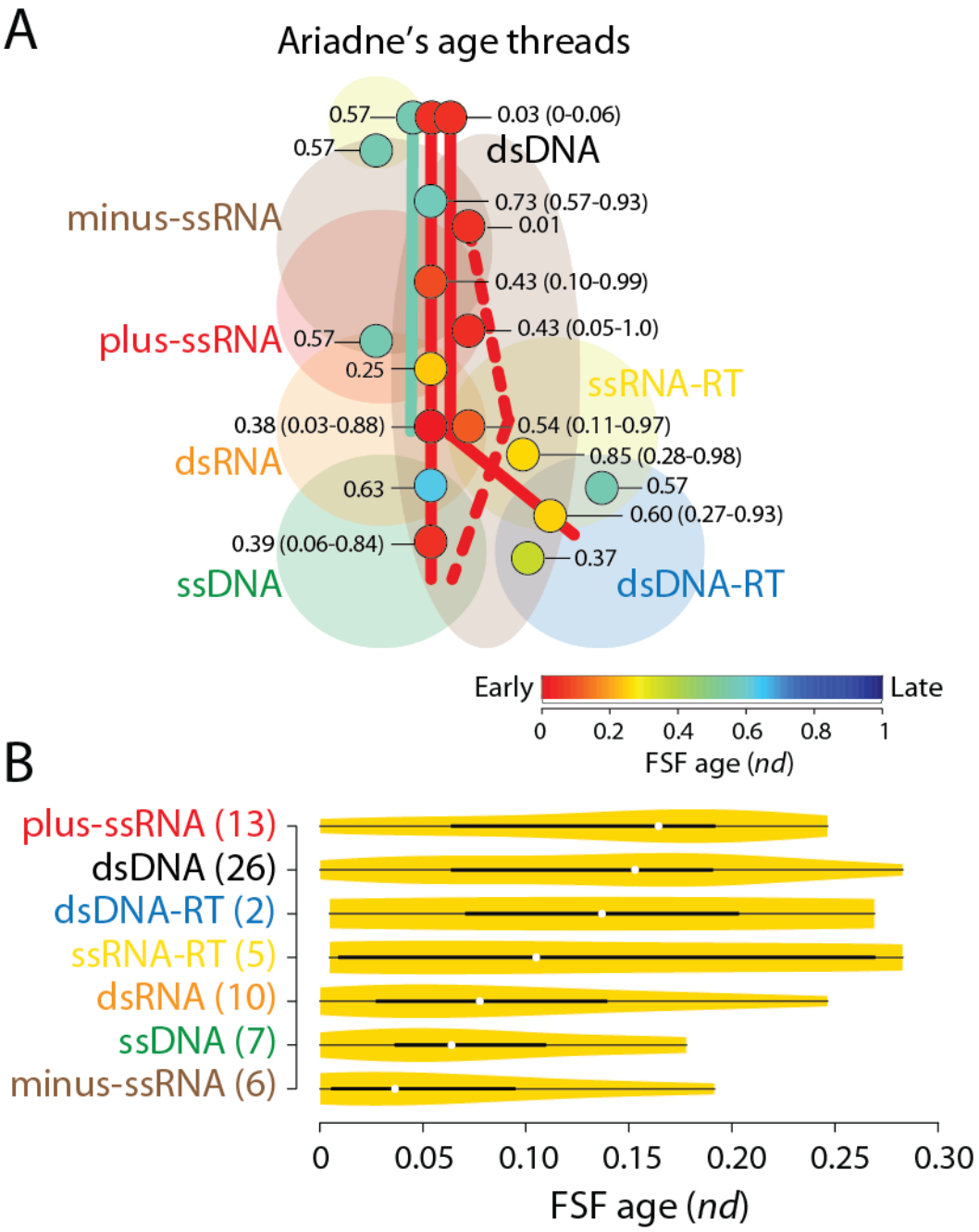

Figure 4.8 The ancient history of RNA viruses. A) The length of Ariadne's threads (colored lines) identifies FSFs that were shared by more than three viral subgroups. Filled circles indicate FSFs shared between two or three viral subgroups. Numbers next to each circle give the mean *nd* of FSFs shared by each combination. Numbers in parenthesis give the range between the most ancient and most recent FSFs that were shared by each combination. **B)** The distribution of the most ancient (*nd* < 0.3) ABEV FSFs in evolutionary timeline (*nd*) for each viral subgroup. Numbers in parenthesis indicate total FSFs in each viral subgroup. White circles indicate group medians. A density trace is plotted symmetrically around the boxplots.

Tables

Table 4.1 VSFs and their distribution within the viral supergroup. FSFs in boldface were novel additions to the list while those in italics were also detected in cells but should be classified as VSFs based on previous knowledge [201] and our analysis.

SCOP Id	SCOP *css*	Venn group	FSF description	Distribution
89428	b.126.1	V	Adsorption protein p2	dsDNA
69070	a.150.1	V	Anti-sigma factor AsiA	dsDNA
90246	h.1.24	V	Head morphogenesis protein gp7	dsDNA
55064	d.58.27	V	Translational regulator protein regA	dsDNA
118208	e.58.1	V	Viral ssDNA binding protein	dsDNA
54957	d.58.8	V	Viral DNA-binding domain	dsDNA
49894	b.28.1	V	Baculovirus p35 protein	dsDNA
82046	b.116.1	V	Viral chemokine binding protein m3	dsDNA
48493	a.120.1	V	gene 59 helicase assembly protein	dsDNA
89433	b.127.1	V	Baseplate structural protein gp8	dsDNA
51289	b.85.5	V	Tlp20, baculovirus telokin-like protein	dsDNA
51332	b.91.1	V	E2 regulatory, transactivation domain	dsDNA
47724	a.54.1	V	Domain of early E2A DNA-binding protein, ADDBP	dsDNA
56548	d.180.1	V	Conserved core of transcriptional regulatory protein vp16	dsDNA
69652	d.199.1	V	DNA-binding C-terminal domain of the transcription factor MotA	dsDNA
57917	g.51.1	V	Zn-binding domains of ADDBP	dsDNA
69908	e.35.1	V	Membrane penetration protein mu1	dsRNA
75347	d.13.2	V	Rotavirus NSP2 fragment, C-terminal domain	dsRNA
48345	a.115.1	V	A virus capsid protein alpha-helical domain	dsRNA
69903	e.34.1	V	NSP3 homodimer	dsRNA
111379	f.47.1	V	VP4 membrane interaction domain	dsRNA
75574	d.216.1	V	Rotavirus NSP2 fragment, N-terminal domain	dsRNA
55671	d.102.1	V	Regulatory factor Nef	ssRNA-RT
47852	a.62.1	V	Hepatitis B viral capsid (hbcag)	dsDNA-RT
56502	d.172.1	V	gp120 core	ssRNA-RT
57647	g.34.1	V	HIV-1 VPU cytoplasmic domain	ssRNA-RT
48045	a.84.1	V	Scaffolding protein gpD of bacteriophage procapsid	ssDNA
88650	b.121.7	V	Satellite viruses	ssDNA
48145	a.95.1	V	Influenza virus matrix protein M1	minus-ssRNA
50012	b.31.1	V	EV matrix protein	minus-ssRNA
75404	d.213.1	V	VSV matrix protein	minus-ssRNA
101089	a.8.5	V	Phosphoprotein XD domain	minus-ssRNA

Table 4.1 (contd.)

SCOP Id	SCOP *css*	Venn group	FSF description	Distribution
143021	d.299.1	V	Ns1 effector domain-like	minus-ssRNA
69922	f.12.1	V	Head and neck region of the ectodomain of NDV fusion glycoprotein	minus-ssRNA
58034	h.1.14	V	Multimerization domain of the phosphoprotein from sendai virus	minus-ssRNA
118173	d.293.1	V	Phosphoprotein M1, C-terminal domain	minus-ssRNA
101156	a.30.3	V	Nonstructural protein ns2, Nep, M1-binding domain	minus-ssRNA
117066	b.1.24	V	Accessory protein X4 (ORF8, ORF7a)	plus-ssRNA
110304	b.148.1	V	Coronavirus RNA-binding domain	plus-ssRNA
143587	d.318.1	V	SARS receptor-binding domain-like	plus-ssRNA
144251	g.87.1	V	Viral leader polypeptide zinc finger	plus-ssRNA
101816	b.140.1	V	Replicase NSP9	plus-ssRNA
103145	d.255.1	V	Tombusvirus P19 core protein, VP19	plus-ssRNA
140367	a.8.9	V	Coronavirus NSP7-like	plus-ssRNA
143076	d.302.1	V	Coronavirus NSP8-like	plus-ssRNA
89043	a.178.1	V	Soluble domain of poliovirus core protein 3a	plus-ssRNA
144246	g.86.1	V	Coronavirus NSP10-like	plus-ssRNA
56983	f.10.1	V	Viral glycoprotein, central and dimerisation domains	plus-ssRNA
141666	b.164.1	V	'SARS ORF9b-like	plus-ssRNA
140506	a.30.8	V	FHV B2 protein-like	plus-ssRNA
101257	a.190.1	V	Flavivirus capsid protein C	plus-ssRNA
158974	**b.170.1**	**V**	**WSSV envelope protein-like**	**dsDNA**
88648	**b.121.6**	**V**	**Group I dsDNA viruses**	**dsDNA**
161240	**g.92.1**	**V**	**T-antigen specific domain-like**	**dsDNA**
160957	**e.69.1**	**V**	**Poly(A) polymerase catalytic subunit-like**	**dsDNA**
56558	**d.182.1**	**V**	**Baseplate structural protein gp11**	**dsDNA**
49889	**b.27.1**	**V**	**Soluble secreted chemokine inhibitor, VCCI**	**dsDNA**
58030	**h.1.13**	**V**	**Rotavirus nonstructural proteins**	**dsRNA**
49818	**b.19.1**	**V**	**Viral protein domain**	**dsRNA, minus-ssRNA, plus-ssRNA**
50176	**b.37.1**	**V**	**N-terminal domains of the minor coat protein g3p**	**ssDNA**
161003	**e.75.1**	**V**	**flu NP-like**	**minus-ssRNA**
160453	**d.361.1**	**V**	**PB2 C-terminal domain-like**	**minus-ssRNA**
160892	**d.378.1**	**V**	**Phosphoprotein oligomerization domain-like**	**minus-ssRNA**
159936	**d.15.14**	**V**	**NSP3A-like**	**plus-ssRNA**
160099	**d.346.1**	**V**	**SARS Nsp1-like**	**plus-ssRNA**

Table 4.1 (contd.)

SCOP Id	SCOP *css*	Venn group	FSF description	Distribution
103068	**d.254.1**	**V**	**Nucleocapsid protein dimerization domain**	**plus-ssRNA**
49749	*b.121.2*	*EV*	*Group II dsDNA viruses VP*	*dsDNA*
103417	*e.48.1*	*EV*	*Major capsid protein VP5*	*dsDNA*
69255	*b.40.8*	*ABEV*	*gp5 N-terminal domain*	*dsDNA*
56826	*e.27.1*	*BV*	*Upper collar protein gp10 (connector protein)*	*dsDNA*
140919	*a.263.1*	*BV*	*DNA terminal protein*	*dsDNA*
101059	*a.159.3*	*BV*	*B-form DNA mimic Ocr*	*dsDNA*
58064	*h.3.1*	*ABEV*	*Influenza hemagglutinin (stalk)*	*dsDNA, minus-ssRNA*
111474	*h.3.3*	*BEV*	*Coronavirus S2 glycoprotein*	*dsDNA, plus-ssRNA*
110132	*b.147.1*	*EV*	*BTV NS2-like ssRNA-binding domain*	*dsRNA*
64465	*d.196.1*	*BV*	*Outer capsid protein sigma 3*	*dsRNA*
55405	*d.85.1*	*EV*	*RNA bacteriophage capsid protein*	*plus-ssRNA*

Table 4.2 Significantly enriched 'biological process' GO terms in VSFs (*FDR* < 10^{-3}).

GO Id	GO term	*Z*-score	*P*-value	*FDR*
GO:0044415	evasion or tolerance of host defenses	16.3	1.62E-06	1.05E-05
GO:0050690	regulation of defense response to virus by virus	16.3	1.62E-06	1.05E-05
GO:0044068	modulation by symbiont of host cellular process	15.45	2.31E-06	1.20E-05
GO:0052572	response to host immune response	14.72	3.18E-06	1.22E-05
GO:0052255	modulation by organism of defense response of other organism involved in symbiotic interaction	14.08	4.24E-06	1.22E-05
GO:0002832	negative regulation of response to biotic stimulus	14.08	4.24E-06	1.22E-05
GO:0051805	evasion or tolerance of immune response of other organism involved in symbiotic interaction	14.08	4.24E-06	1.22E-05
GO:0019048	modulation by virus of host morphology or physiology	13.52	5.50E-06	1.43E-05

Table 4.3 FSFs involved in capsid/coat assembly processes in viruses. The *f*-value in cells indicates total number of cellular proteomes (Archaea, Bacteria, and Eukarya combined) encoding an FSF divided by the total number of proteomes.

SCOP Id	SCOP *css*	FSF description	Viral lineage	*f*-value in cells
82856	e.42.1	L-A virus major coat protein	BTV-like	0.25
56831	e.28.1	Reovirus inner layer core protein p3	BTV-like	0.19
48345	a.115.1	A virus capsid protein alpha-helical domain	BTV-like	0.00
56563	d.183.1	Major capsid protein gp5	HK97-like	23.52
103417	e.48.1	Major capsid protein VP5	HK97-like	0.06
88633	b.121.4	Positive stranded ssRNA viruses	Picornavirus-like	3.64
88645	b.121.5	ssDNA viruses	Picornavirus-like	0.99
88650	b.121.7	Satellite viruses	Picornavirus-like	0.00
88648	b.121.6	Group I dsDNA viruses	Picornavirus-like	0.00
49749	b.121.2	Group II dsDNA viruses VP	PRD1/Adenovirus-like	0.31
47353	a.28.3	Retrovirus capsid dimerization domain-like	Other/Unclassified	4.07
47943	a.73.1	Retrovirus capsid protein, N-terminal core domain	Other/Unclassified	1.23
47195	a.24.5	TMV-like viral coat proteins	Other/Unclassified	0.99
57987	h.1.4	Inovirus (filamentous phage) major coat protein	Other/Unclassified	0.68
51274	b.85.2	Head decoration protein D (gpD, major capsid protein D)	Other/Unclassified	0.49
64465	d.196.1	Outer capsid protein sigma 3	Other/Unclassified	0.06
55405	d.85.1	RNA bacteriophage capsid protein	Other/Unclassified	0.06
48045	a.84.1	Scaffolding protein gpD of bacteriophage procapsid	Other/Unclassified	0.00
47852	a.62.1	Hepatitis B viral capsid (hbcag)	Other/Unclassified	0.00
101257	a.190.1	Flavivirus capsid protein C	Other/Unclassified	0.00
50176	b.37.1	N-terminal domains of the minor coat protein g3p	Other/Unclassified	0.00
103068	d.254.1	Nucleocapsid protein dimerization domain	Other/Unclassified	0.00

Table 4.4 Descriptive statistics on the results of HMM assignments in each virus subgroup. Proteomic coverage is the number of proteins with FSFs assignments divided by the total number of proteins in a proteome and multiplied by 100. This value exceeds 100% for plus-ssRNA and ssRNA-RT viruses as they often encode single polyproteins that are later processed into individual subunits.

Type	***N***	***N'***	***N''***	***N'''***	***M***	***M'***	***M''***	***M'''***
dsDNA	1649	170602	44723	29552	103.46	27.12	17.92	26.21
ssDNA	534	3148	784	691	5.90	1.47	1.29	24.90
dsRNA	166	753	347	328.	4.54	2.09	1.98	46.08
plus-ssRNA	880	3318	5653	3196	3.77	6.42	3.63	170.37
minus-ssRNA	111	751	301	279	6.77	2.71	2.51	40.08
ssRNA-RT	56	237	681	396	4.23	12.16	7.07	287.34
dsDNA-RT	64	301	185	177	4.70	2.89	2.77	61.46
Supergroup	3460	179110	52674	34619	51.77	15.22	10.01	29.41

N total number of proteomes in each viral subgroup
N' total number of proteins from all proteomes in each viral subgroup
N'' total number of FSFs detected in the entire proteomic set of each viral replicon
N''' total number of unique FSFs detected in the entire proteomic set of each viral replicon
M mean length of proteome in each viral subgroup
M' mean number of total FSFs assigned for all proteomes in a subgroup
M'' mean number of unique FSFs detected in all proteomes in a subgroup
M''' mean proteomic coverage

Table 4.5 FSFs shared between different viral subgroups.

SCOP Id	SCOP *css*	FSF description	Distribution
56672	e.8.1	DNA/RNA polymerases	dsDNA, dsRNA, dsDNA-RT, ssRNA-RT, minus-ssRNA, plus-ssRNA
52540	c.37.1	P-loop containing nucleoside triphosphate hydrolases	dsDNA, dsRNA, ssDNA, minus-ssRNA, plus-ssRNA
53335	c.66.1	S-adenosyl-L-methionine-dependent methyltransferases	dsDNA, dsRNA, ssDNA, minus-ssRNA, plus-ssRNA
53098	c.55.3	Ribonuclease H-like	dsDNA, ssRNA-RT, ssDNA, minus-ssRNA
88633	b.121.4	Positive stranded ssRNA viruses	dsDNA, dsRNA, minus-ssRNA, plus-ssRNA
57850	g.44.1	RING/U-box	dsDNA, minus-ssRNA, plus-ssRNA
51283	b.85.4	dUTPase-like	dsDNA, dsDNA-RT, ssRNA-RT
56112	d.144.1	Protein kinase-like (PK-like)	dsDNA, dsRNA, ssRNA-RT
54768	d.50.1	dsRNA-binding domain-like	dsDNA, dsRNA, plus-ssRNA
54001	d.3.1	Cysteine proteinases	dsDNA, minus-ssRNA, plus-ssRNA
52266	c.23.10	SGNH hydrolase	dsDNA, minus-ssRNA, plus-ssRNA
58100	h.4.4	Bacterial hemolysins	dsDNA, dsRNA, ssDNA
49818	b.19.1	Viral protein domain	dsRNA, minus-ssRNA, plus-ssRNA
57756	g.40.1	Retrovirus zinc finger-like domains	dsDNA, dsDNA-RT, ssRNA-RT
50044	b.34.2	SH3-domain	dsDNA, dsRNA, ssRNA-RT
57924	g.52.1	Inhibitor of apoptosis (IAP) repeat	dsDNA, plus-ssRNA
50249	b.40.4	Nucleic acid-binding proteins	dsDNA, ssDNA
53041	c.53.1	Resolvase-like	dsDNA, ssDNA
55550	d.93.1	SH2 domain	dsDNA, ssRNA-RT
55464	d.89.1	Origin of replication-binding domain, RBD-like	dsDNA, ssDNA
56399	d.166.1	ADP-ribosylation	dsDNA, ssDNA
100920	b.130.1	Heat shock protein 70kD (HSP70), peptide-binding domain	dsDNA, plus-ssRNA
47413	a.35.1	lambda repressor-like DNA-binding domains	dsDNA, ssDNA
69065	a.149.1	RNase III domain-like	dsDNA, plus-ssRNA
46785	a.4.5	Winged helix DNA-binding domain	dsDNA, ssDNA
53448	c.68.1	Nucleotide-diphospho-sugar transferases	dsDNA, dsRNA
57997	h.1.5	Tropomyosin	dsDNA, dsRNA
54236	d.15.1	Ubiquitin-like	dsDNA, ssRNA-RT
47954	a.74.1	Cyclin-like	dsDNA, ssRNA-RT
90229	g.66.1	CCCH zinc finger	dsDNA, minus-ssRNA
103657	a.238.1	BAR/IMD domain-like	dsDNA, ssRNA-RT
53067	c.55.1	Actin-like ATPase domain	dsDNA, plus-ssRNA
47794	a.60.4	Rad51 N-terminal domain-like	dsDNA, ssDNA
143990	d.336.1	YbiA-like	dsDNA, plus-ssRNA
55811	d.113.1	Nudix	dsDNA, dsRNA

Table 4.5 (contd.)

SCOP Id	SCOP *css*	FSF description	Distribution
51197	b.82.2	Clavaminate synthase-like	dsDNA, plus-ssRNA
53756	c.87.1	UDP-Glycosyltransferase/glycogen phosphorylase	dsDNA, dsRNA
81665	f.33.1	Calcium ATPase, transmembrane domain M	dsDNA, plus-ssRNA
52949	c.50.1	Macro domain-like	dsDNA, plus-ssRNA
53955	d.2.1	Lysozyme-like	dsDNA, dsRNA
49899	b.29.1	Concanavalin A-like lectins/glucanases	dsDNA, dsRNA
48371	a.118.1	ARM repeat	dsDNA, plus-ssRNA
51126	b.80.1	Pectin lyase-like	dsDNA, plus-ssRNA
47598	a.43.1	Ribbon-helix-helix	dsDNA, ssDNA
50494	b.47.1	Trypsin-like serine proteases	dsDNA, plus-ssRNA
55144	d.61.1	LigT-like	dsDNA, plus-ssRNA
81296	b.1.18	E set domains	dsDNA, plus-ssRNA
161008	e.76.1	Viral glycoprotein ectodomain-like	dsDNA, minus-ssRNA
90257	h.1.26	Myosin rod fragments	dsDNA, dsRNA
57501	g.17.1	Cystine-knot cytokines	dsDNA, ssRNA-RT
54117	d.9.1	Interleukin 8-like chemokines	dsDNA, dsRNA
58069	h.3.2	Virus ectodomain	ssRNA-RT, minus-ssRNA
50630	b.50.1	Acid proteases	dsDNA-RT, ssRNA-RT
47459	a.38.1	HLH, helix-loop-helix DNA-binding domain	dsDNA, ssRNA-RT
50939	b.68.1	Sialidases	dsDNA, minus-ssRNA
55166	d.65.1	Hedgehog/DD-peptidase	dsDNA, ssDNA
51225	b.83.1	Fibre shaft of virus attachment proteins	dsDNA, dsRNA
49835	b.21.1	Virus attachment protein globular domain	dsDNA, dsRNA
111474	h.3.3	Coronavirus S2 glycoprotein	dsDNA, plus-ssRNA
55658	d.100.1	L9 N-domain-like	dsDNA, dsDNA-RT
55895	d.124.1	Ribonuclease Rh-like	dsDNA, plus-ssRNA
52972	c.51.4	ITPase-like	dsDNA, plus-ssRNA
57959	h.1.3	Leucine zipper domain	dsDNA, ssRNA-RT
50203	b.40.2	Bacterial enterotoxins	dsDNA, ssDNA
48208	a.102.1	Six-hairpin glycosidases	dsDNA, ssDNA
50022	b.33.1	ISP domain	dsDNA, ssRNA-RT
58064	h.3.1	Influenza hemagglutinin (stalk)	dsDNA, minus-ssRNA

Table 4.6 Significantly enriched 'biological process' GO terms in EV FSFs (*FDR* < 10^{-2}). No terms were enriched in either AV or BV FSFs.

GO Id	GO description	*Z*-score	*P*-value	*FDR*
GO:0050918	positive chemotaxis	10.87	4.26E-07	6.06E-05
GO:0010634	positive regulation of epithelial cell migration	9.62	4.41E-07	6.06E-05
GO:0001569	patterning of blood vessels	7.7	5.70E-05	8.71E-04
GO:0050921	positive regulation of chemotaxis	6.45	2.22E-04	1.97E-03
GO:0046888	negative regulation of hormone secretion	7.3	3.23E-04	2.34E-03
GO:0010594	regulation of endothelial cell migration	5.79	4.76E-04	3.19E-03
GO:0060425	lung morphogenesis	6.34	7.54E-04	4.23E-03
GO:0050829	defense response to Gram-negative bacterium	6.34	7.54E-04	4.23E-03
GO:0048640	negative regulation of developmental growth	5.43	7.31E-04	4.23E-03
GO:0003156	regulation of organ formation	5.97	1.06E-03	5.29E-03
GO:0010464	regulation of mesenchymal cell proliferation	5.37	1.89E-03	7.52E-03
GO:0002062	chondrocyte differentiation	5.37	1.89E-03	7.52E-03
GO:0048483	autonomic nervous system development	5.37	1.89E-03	7.52E-03
GO:0045766	positive regulation of angiogenesis	5.12	2.42E-03	8.87E-03
GO:0090100	positive regulation of transmembrane receptor protein serine/threonine kinase signaling pathway	5.12	2.41E-03	8.86E-03

CHAPTER 5: THE DISTRIBUTION AND IMPACT OF VIRAL LINEAGES IN DOMAINS OF LIFE[5]

Viruses impact our economy, medicine and agriculture due to their infectious nature. Viral infections transform the host cell into a 'virocell' that no longer divides by binary fission but produces more viral particles or a 'ribovirocell' in which the viral and cellular genomes coexist, the cell still dividing while producing virions [240,241]. Here, we address the impact of viral infections on the evolution of cells exhaustively study viral host preferences. Specifically, we consider that gain and loss of viral lineages often leads to divergent evolutionary trends even in closely related species. We emphasize that no evolutionary theory could be complete without accounting for the viral world and that viruses are responsible for ongoing adaptations in the cellular domains (see also [321,322]).

The distribution of the association of viral replicon types with cells is extremely biased. For example, RNA viruses are completely absent in Archaea and are rare in Bacteria. In comparison, vertebrates host numerous RNA and retroviruses. Surprisingly, dsDNA viruses are rare in plants while dsRNA viruses are abundant in fungi. Similarly, retroviruses are integrated into the genomes of multicellular eukaryotes but are completely absent in the microbial genomes. In other words, specific relationships exist between the type of viral replicon and the host range. Viruses with a particular replicon may infect one group of organisms but may not replicate in another. Big jumps of viruses from one cellular lineage to another have been observed within the eukaryotic 'division' such as animals (opisthokonts) and plants (viridiplantae), when a virus adapts to an established consortium of ecological partners. The same virus can sometimes infect both plant and animal cells when these are linked by their mode of life. One example is the *Fiji disease virus* (*Reoviridae*) that can replicate in both its insect vector (Delphacidae) and flowering plants [225]. However, no modern virus is known to cross the barrier between domains. Therefore, while viruses may be able to jump hosts over short evolutionary time spans, crossing domain boundaries is less likely and not expected to compromise our inferences.

To obtain a quantitative view of viral diversity and its distribution among cellular domains, we extracted genome data from the Viral Genomes Resource at NCBI [254]. This

[5]This chapter has been published as manuscript in *Frontiers in Mircrobiology* (see [212]). The final publication is available at http://journal.frontiersin.org/Journal/10.3389/fmicb.2014.00194/full. Authors retain the rights to reprint.

resource provides accurate, manually curated information about sequenced viral genomes that is minimally redundant. Generally, one sequenced genome portrays many isolates/strains of the same virus. Specifically, we investigated the host preferences for viruses with different replication strategies (Figure 5.1A) and contrasted virion morphologies (borrowed from ViralZone; [320]) of virus families infecting different domain groups (Figure 5.1B).

A 'birds-eye' view of the distribution of viruses among hosts revealed that only 63 were exclusive to the archaeal domain (hereinafter referred to as archaeoviruses) (Figure 5.1A). In comparison, 1,251 bacterial (bacterioviruses, formerly bacteriophages) and 2,321 eukaryal viruses (eukaryoviruses) were identified. The low number of archaeoviruses is clearly due to a sampling bias (e.g. the low number of archaeal species screened for the presence of viral infection) since it has been shown that four different viruses can infect a single archaeal species (i.e. *Aeropyrum pernix*), each from a different family [319,323,324]. Despite their low number, archaeoviruses exhibit greater virion morphotype diversity compared to bacterioviruses [e.g. 4 unique virion morphotypes vs. none (Figure 5.1B); see also [210]. In comparison, bacterial organisms host a vast number of described DNA viruses (1,178 out of total 1,760) but display very little family and morphotype diversity. In fact, 95% of the dsDNA bacterioviruses belong to just one order (*Caudovirales*) and only three families (*Myoviridae*, *Podoviridae*, and *Siphoviridae*). Moreover, only 9 virion morphologies have been observed in bacterioviruses (compared to 16 in Archaea) [210]. One explanation for the low diversity of bacterioviruses could be the invention of peptidoglycan-containing cell wall in Bacteria. The inability to traverse this barrier likely resulted in loss of many viral lineages in Bacteria [321,325]. Taken together, these observations suggest that Archaea are likely infected by a greater number of viral lineages than Bacteria. This is showcased by their virion morphologies diversity (Figure 5.1B) [210,211], which is expected to grow with improvements in our ability to isolate viruses from atypical habitats.

Interestingly, all archaeoviruses possess DNA replicons but no RNA genomes. The complete absence of RNA viruses in Archaea can be linked to high temperature RNA instability [70]. We speculate that escape from RNA viruses could be one major trigger for the evolution of modern Archaea [70]. Thus, loss of RNA viral lineages likely initiated archaeal migration to the harsh environments. One recent study reported the isolation of ssRNA(+) viruses from an archaea-rich community in a hot, acidic spring of Yellowstone National Park [326]. However,

their host tropism could not be established with confidence. Finally, four ssDNA viruses were recently isolated from Archaea [318,319,327]. Of these, *Aeropyrum* coil-shaped virus (*Spiraviridae*) is the largest known ssDNA virus and displays unique coil-shaped virion morphology [319].

Bacterioviruses are remarkably successful in Bacteria and are highly abundant. Their virions outnumber Bacteria in oceans, balance microbial populations in the marine communities, and regulate biogeochemical cycles [208,209,328]. Among the dsDNA bacterioviruses, tailed-bacteriophages exhibit extensive similarities with archaeal *Caudovirales*, suggesting that they form a monophyletic group [329]. Archaeal and bacterial *Caudovirales* have indeed been grouped in a single major evolutionary lineage, together with *Herpesviridae*. All of these viruses share the same Hong Kong fold (HK97) in their major capsid proteins and homologous packaging ATPases [330]. Notably, it has been found recently that the capsid of *Herpesviridae* exhibits a small tail similar to those of *Podoviridae* [278]. These data suggest that viruses of the HK97-like lineage are very ancient and originated (most likely) prior to the last common ancestor of cells. Another example of viral lineage shared by the three domains is the so-called 'PRD1/Adenovirus lineage' of dsDNA viruses characterized by a major capsid protein containing the double-jelly roll fold and a common packaging ATPase [227]. In comparison, ssDNA bacterioviruses are not as successful in Bacteria and correspond to two major families, *Inoviridae* and *Microviridae* (smallest genomes among DNA viruses; [331]). Viruses in this group replicate by converting their single-stranded DNA genome into a double-stranded intermediate form engineered by host polymerase. These viruses lack their own polymerase and share this property with the ssDNA viruses of Archaea and Eukarya.

In contrast to DNA viruses, RNA viruses are not as successful in Bacteria. Only, 5 dsRNA, and 11 ssRNA(+) bacterioviruses could be identified. In turn, none of the ssRNA(-) and retrotranscribing viruses associated with bacterial hosts. Among the RNA bacterioviruses, dsRNA viruses (*Cystoviridae*) encode segmented genomes and infect mostly Pseudomonas species [332]. Interestingly, *Cystoviridae* closely resembles eukaryal dsRNA viruses (i.e. *Reoviridae* and *Totiviridae*) in terms of life cycle and homologous RNA-dependent-RNA-polymerase gene sequences (a viral hallmark) [333]. Unlike Archaea, Bacteria are also infected by ssRNA(+) viruses (*Leviviridae*). These viruses are amongst the simplest and smallest known viruses, and historically yielded useful insights into mRNA function [334]. Because RNA viruses

(ssRNA and dsRNA) infect both Bacteria and Eukarya, their ancestors likely originated from a putative ancient world of cells with RNA genomes and RNA viruses [194]. This points to the ancient existence of RNA viruses and suggests their loss from Archaea (since loss in one domain is more likely than the independent gain in two!). The instability of RNA at high temperatures supports this hypothesis, since it is likely that the last common ancestor of Archaea was a hyperthermophile [163].

Viruses with all possible types of replicons infect eukaryal organisms. RNA viruses are predominant and cover the entire taxonomic range within Eukarya (Figure 5.1A). Eukaryoviruses also exhibit many unique virion morphotypes not observed in the prokaryotic viruses and are unequally distributed in the major eukaryal groups (Figure 5.1). For example, dsDNA viruses are completely absent in fungi and are rare in plants (i.e. only found in green algae). This suggests that these groups have evolved sophisticated mechanisms to eliminate dsDNA viral infections. A good candidate is the cell wall structure found in plants, fungi, and algae. Differences in cell wall composition and rigidity greatly limit means of viral entry into the cell and serve as barriers to viral infections [335]. However, loss of one viral lineage is apparently offset by the gain of other lineages. This is evident from the high RNA virus distribution among plants and fungi. The origin of the diversity and abundance of RNA viruses in eukaryotes but their near absence in prokaryotes is particularly puzzling [207]. For example, ssRNA(-) and retroviruses are highly successful in vertebrates. At first glance, it seems that organism complexity is proportional to the variety of viral infections. For instance, metazoa are infected by a host of retroviruses. Retroviruses can integrate their genomes into host DNA and thus alter gene expression patterns and trigger genomic rearrangements [336]. These activities can lead to production of novel genes and advanced machineries [70]. In fact, telomerase enzymes are homologous to retroviral proteins and neocentromeres are formed by epigenetic regulation of transposable elements [337,338], both likely transferred from viruses to host cells much earlier in evolution. This argument is further supported by the absence of RNA and retroviruses from unicellular eukaryotes such as yeast, which resemble a prokaryotic lifestyle [70]. Thus, co-evolution between viruses and their hosts may have led to organism complexity in the eukaryotic domain.

The diversity of eukaryotic viruses is intriguing, both in terms of genome structure and virion morphology (see Figure 5.1B). In particular, retrotranscribing, ssRNA(-), and many DNA

virus families are only present in eukaryotes. Surprisingly, although Archaea and Eukaryotes are very similar in term of their basic molecular biology, there are no viral lineages specific for these two domains [70]. Virions with rod-shaped morphology are up to now specific for Archaea and Eukarya (Figure 5.1B), but they harbor DNA and RNA genomes, respectively, and it is unclear if their major coat proteins are evolutionary related [339]. The same is probably also true for bacilliform viruses. Notably, the diversity and specificity of eukaryoviruses is difficult to reconcile with the archaeon-bacterium fusion scenarios for the origin of eukaryotes (e.g. [67]), as recently argued [70].

To conclude, the distribution of viral lineages follows an ancient, highly dynamic and ongoing process that impacts the evolution of organisms. New viral lineages often arise from existing ones and may cross species barriers to infect new hosts (e.g. parvoviruses; [340], putting enormous evolutionary pressure on cellular organisms and prompting them to unfold molecular and cellular innovation [200] in the search of either simplicity or complexity.

Figures

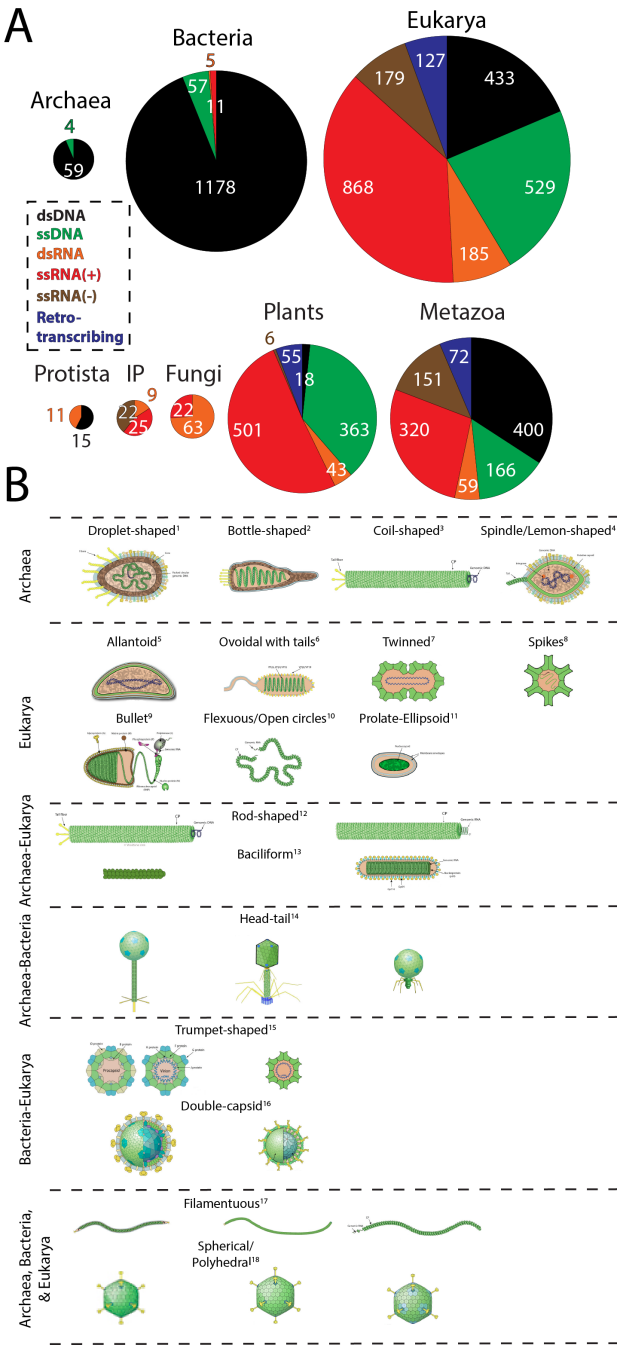

Figure 5.1 The abundance and diversity of viral lineages in the domains of life. A) Pie-charts describe the abundance of dsDNA, ssDNA, dsRNA, ssRNA(+), ssRNA(-), and retrotranscribing viruses in Archaea, Bacteria, and Eukarya, and within the major eukaryal divisions. Genome data from 3,660 completely sequenced viral genomes corresponding to 1,671 dsDNA, 610 ssDNA, 883 ssRNA(+), 179 ssRNA(-), 190 dsRNA, and 127 retrotranscribing viruses were retrieved from the Viral Genomes Resource (April 2014). Additionally, two ssDNA archaeal viruses were identified from the literature [318,319]. Viruses that were unassigned to any order, genera, or species and unclassified viruses were excluded from sampling. Viruses were broadly classified according to host preferences into the following categories: Archaea, Bacteria, Protista (animal-like protists and brown algae),

Invertebrates and plants (IP); Fungi (all fungi and fungi-like protists); Plants (all plants, green algae, and diatoms), and Metazoa (vertebrates, invertebrates, and human). Host information was available for roughly 99% (3,633) of the sampled viruses. Pie-charts are proportional to the size of each distribution. **B)** Virion morphotypes that are specific to a domain or are shared between domains are displayed. Virion pictures were borrowed from the ViralZone web-resource [320] and from [210]. A keyword-based search was performed on text data to assign the most general morphotypes (e.g. rod-shaped, spherical, droplet-shaped, etc.) to all viruses. More than one viridae with same morphotype is possible but not made explicit. The diagram does not always imply evolutionary relationship between viruses harboring common morphology. For example, archaeal and eukaryal rod-shaped viruses are probably not evolutionarily related [339]. Well-studied exceptions are of head-tail *Caudovirales* harboring the HK97 capsid fold and of polyhedral viruses harboring the 'double jelly-roll' fold [227]. [1]*Guttaviridae*; [2]*Ampullavirus*; [3]*Spiraviridae*; [4]*Fuselloviridae*; [5]*Ascoviridae*; [6]*Nimaviridae*; [7]*Geminiviridae*; [8]*Astroviridae*; [9]*Rhabdoviridae*; [10]*Ophioviridae*; [11]*Polydnaviridae*; (left to right) [12]*Rudiviridae* [Archaea]; *Virgaviridae* [Eukarya]; [13]*Clavaviridae* [Archaea]; *Roniviridae* [Eukarya]; [14]*Siphoviridae, Myoviridae,* and *Podoviridae* [Archaea and Bacteria]; [15]*Microviridae* [Bacteria], *Circoviridae* [Eukarya]; [16]*Cystoviridae* [Bacteria], *Reoviridae* [Eukarya]; [17]*Lipothrixiviridae* [Archaea], *Inoviridae* [Bacteria], *Potyviridae* [Eukarya]; [18]*Sulfolobus* turreted icosahedral virus [Archaea], *Tectiviridae* [Bacteria], *Adenoviridae* [Eukarya].

CHAPTER 6: UNTANGLING THE ORIGIN OF VIRUSES AND THEIR IMPACT ON CELLULAR EVOLUTION[6]

The fact that viruses infecting distantly related hosts share specific proteins (e.g. major capsid proteins and packaging ATPases) that lack homologs in the proteomes of cellular organisms suggests that the viral mode of life originated very early in evolution. This argument is further supported by the great diversity seen in viral replication strategies (e.g. DNA, RNA, and retrotranscribing viruses) and virion morphologies [210,212]. However, all modern viruses require an intracellular environment for viral protein synthesis. Therefore, it is difficult to predict the nature of ancient 'viruses' and how they survived prior to the appearance of modern cells. One explanation is that viruses originated from ancient cells by gene loss or reductive evolution [219,220,267]. Recently, this idea has become popular with the discovery of many 'giant' viruses (e.g. mimiviruses, pandoraviruses, megaviruses, and pithoviruses) that surpass many cellular parasitic species in particle and genome size [214-217]. Remarkably, giant viruses also encode some key proteins involved in protein translation (e.g. aminoacyl-tRNA synthetases) [341], suggesting that perhaps a full or rudimentary translation apparatus was once present in the ancestral virus [14,217] (for an opposite view see [257]).

The reductive scenario of viral origins is also supported by the observation that the tendency to lose or replace genes inside a cellular host is a recurring phenomenon in cellular parasites. For example, many bacterial species possess highly reduced genomes and survive as endosymbionts of other species [268]. An extreme case of reductive evolution is the mitochondrion, which is the result of substantial gene loss in ancestral α-proteobacteria and permanent integration into the cellular makeup of its primordial host [342,343]. However, a similar phenomenon for viral evolution is rarely invoked (except see [219,220,267]) despite the fact that all viruses are obligate intracellular parasites and must infect or become part of cellular genomes (e.g. endogenous retroviruses) to reproduce. In turn, viral evolution is largely credited to gene uptake from cells via horizontal gene transfer (HGT) [237] mainly because some key viral proteins show high sequence-similarities with proteins in cellular organisms. However, sequence-based evolutionary studies have largely failed to paint the complete picture of viral

[6]This chapter has been submitted for publication to *Annals of the New York Academy of Science* and is currently under review.

evolution. Most importantly, a large fraction of viral proteomes does not show significant sequence similarity to any of the known cellular proteins (e.g. see [344]). Nevertheless, the use of molecular structure (and functions) has recently become popular in the evolutionary studies of cells and viruses [14,20,256,301]. Protein domain and tRNA secondary structures are typically less prone to the effect of mutations and are evolutionarily more conserved than nucleotide or protein sequences [37]. Recently, the usage and distribution of protein domains defined at the SCOP fold superfamily (FSF) level [33,34] was compared across proteomes of large dsDNA viruses and cellular organisms [14]. The findings confirmed an early origin of large DNA viruses from ancient cells and subsequent adaptation to parasitism [195]. However, a similar origin may be difficult to fathom for other viruses, especially positive-sense RNA viruses, as they encode very few proteins and resemble the 'simplified' forms of cellular mRNAs. Thus, they could in fact be products of modern cells that 'escaped' cellular control and became infectious. In other words, it is possible that different groups of viruses appeared at multiple times in evolution and via distinct evolutionary mechanisms.

These considerations beg an important question. Which hypothesis completely and adequately explains the origin of viruses and carries maximum explanatory power? Because sequence-based studies cannot completely capture the entire diversity of the virosphere, we suggest focusing on atypical features to objectively answer this important question. In this opinion article, we propose three promising lines of research that could help test different scenarios of viral origin and evolution: (i) virus-host interactions, (ii) morphological similarities in virion particles, and (iii) structural data from the evolutionary studies of protein domains and tRNA molecules. We first review the most recent data pertaining to each of the three lines of research and then discuss how they could be used in improving our understanding of viruses and their evolution. We attempt to take a balanced approach and highlight both the *pros* and *cons* of each of the three main research directions.

What does virus-host preferences tell us about viral evolution?

All viruses must replicate inside a cellular host. Sometimes, this results in an infection of the host cell while in other cases both viral and cellular genomes coexist [241]. We argue that the strong cellular dependency of viruses indicates a long-term symbiotic-like interaction that has greatly influenced the evolution and makeup of modern cells. There are significant biases in the

host range of viruses harboring different replicon types that could be very informative in inferring the appearance order of different viral replicons. For example, the giant DNA viruses infect members of several major eukaryotic divisions (e.g. Opisthokonts, Amoebozoa, Archaeplastida), suggesting that they were present very early in eukaryotic evolution. Similarly, most species of RNA viruses are specific to eukaryotes [212]. These include the minus-sense RNA, retroviruses and pararetroviruses. In contrast, they have not (*yet*) been detected in microbial species, including akaryotes (i.e. cell without nucleus; previously prokaryote) [345].

At first glance, this data suggest that RNA viruses originated late in evolution and could have evolved from RNA families within the eukaryotic cells (i.e. favoring the 'escape' hypothesis for viral evolution). In turn, DNA viruses may be very ancient as they cover a broad spectrum of hosts on the tree of life (ToL). However and despite their appeal, interpretations like these should be taken with caution. One source of error could be the ascertainment bias. For example, there is strong motivation to study/isolate human (and vertebrate/plant) viruses due to obvious medical and economical reasons. Thus, many akaryotic viruses have likely escaped detection and could change the overall picture. However, if we restrict ourselves to available and most recent data on viruses and their hosts (taken from NCBI viral genomes resource [254]), it is clear that relatively few RNA viral families infect akaryotic hosts and a greater number infect the eukaryotic organisms (argued in detail in [212]). This argument makes sense since it is likely that the inability to infect a particular host did not occur by chance but rather involved major evolutionary innovations that resulted in the 'loss' of viral lineages in potential host organisms.

We speculate that loss of RNA viruses in both Archaea and Bacteria may be a more parsimonious explanation than the late origin of RNA viruses. For example, most members of the archaeal domain have been isolated from harsh environments favoring very high temperatures and saline conditions. Given the instability of RNA at excessive temperatures, it is reasonable to think that the emerging archaeal cells adapted to extreme environments to escape from invading RNA viruses [70]. In turn, Bacteria evolved peptidoglycan-containing cell walls that are difficult to penetrate [321,325] except by most head-tailed viruses belonging to order *Caudovirales* that often mediate genetic exchange between bacterial species and likely help drive bacterial evolution [212]. Thus selection of viruses could be one significant factor in bacterial evolution. In contrast, eukaryotes are infected by a large number of viruses from all replicon types and likely benefited from this evolutionary 'arms race' [70,200,202]. This phenomenon

may still be occurring as modern cells are constantly challenged by a large number of emerging novel viral lineages (e.g. Ebola, MERS, and Influenza viruses). The most anticipated outcome of the constant battle between modern viruses and cells is to invoke novel mechanisms of escape from viral infections (i.e. by loss of viral replicons). This ongoing battle either directs the evolution of organisms towards simplicity (as in akaryotes) or more complexity (as in eukaryotes) [212].

To summarize, the early origins of RNA viruses is incompatible with the current distribution of viral replicons in host organisms unless one invokes early loss of RNA replicons from akaryotic microbes. We propose that the 'loss' scenario may be more likely given the physiological and molecular makeup of modern cells and because RNA viruses are smaller in size and are mutation-prone, which is consistent with the perceived genetic system of the last universal cellular ancestor (LUCELLA) [346]. Further support of this argument comes from the observation that there is a path from RNA to DNA viruses via the retrotranscribing viruses [194]. Thus, if carefully interpreted and assuming that sampling biases will not make a drastic difference, available data on virus-host preferences give significant clues regarding the origin and evolution of different viral families.

The many known virion morphotypes likely originated from a rather small number of structural designs

When we look at different virion shapes associated with viruses infecting the three cellular domains, Archaea, Bacteria, and Eukarya, we note that the more complex morphotypes were restricted to either archaeal (archaeoviruses) or eukaryal viruses (eukaryoviruses) (Figure 6.1). In turn, no unique morphology was detected in bacterial viruses (bacterioviruses). Thus, archaeoviruses are more diverse in morphotype number than bacterioviruses [210]. In contrast, bacterioviruses are mostly restricted to the head-tailed *Caudovirales* and lack morphotype diversity.

From a tensegrity point-of-view, the 'spherical' and 'filamentous' morphotypes shared by the viruses of three domains are the simplest virion architectures. There seems to be an interesting pattern stemming from these common designs. The 'filamentous' spread their design towards Archaea and the 'spherical/polyhedral' towards Eukarya. The common 'head-tail' morphology in Archaea and Bacteria likely combines the two common designs into one. In other

words, a common set of simpler forms gives rise to a multitude of complex structures. However, morphological similarities may not always be a result of vertical evolution and need to be supported with other molecular data. Interestingly, the 'head-tail' *Caudovirales* are united by the presence of the 'HK 97' capsid protein fold and the 'spherical/polyhedral' viruses by the 'double jelly roll' protein fold [227]. Member viruses of these two morphotypes infect organisms from all cellular domains and likely evolved prior to LUCELLA. Thus, structural similarities between apparently very distantly related viruses also hold strong clues regarding the evolution of viruses. In short, the many different virion shapes observed in modern day viruses may be traced back to a small number of structural designs (spherical and filamentous that could have appeared independently in evolution) that were present in ancestral viruses.

Protein domain structures tell a lot about the evolutionary history of cells and viruses

SCOP defines FSFs to include protein domains that are evolutionarily related. At the FSF level of SCOP hierarchy, protein domains exhibit very little sequence identity (as low as <15%) but share structural and biochemical properties that are indicative of common origin [33,34]. Because molecular structure is relatively more robust against mutations that change the nucleotide or protein sequence [37], FSF domains provide the ideal characters to study long-term evolution [201]. For example, the capsid proteins of Eel picornavirus 1 and Rabbit hemorrhagic disease virus exhibit very little sequence identity (not more than 28%) but can still be recognized by SUPERFAMILY HMMs (see [40,41] for FSF assignment protocol). This approach has been used previously in a large number of studies involving both viral and cellular proteomes [14,20,27,201].

Another advantage of using FSF domains is to make reliable inferences regarding the age of modern proteomes. FSF evolution follows a clock-like behavior that has been linked to geological record [51]. Thus, FSFs provide reliable estimates regarding the onset of key events in the evolutionary history of organisms. Here, we investigate the spread of 1,993 FSFs that were detected in 4,211 cellular and viral proteomes along an evolutionary timescale, with time provided as a *node distance* (*nd*). *nd* is a substitute for the true age of an FSF and was calculated from a phylogenetic tree of domains [20,29]. The phylogenetic model assumes that FSFs that are more *abundant* and *widespread* should be more ancient relative to those with low abundance and narrow spread. For example, the P-loop containing NTP hydrolase domains and some other

metabolic folds that are universal among cellular organisms likely appeared first in evolution. In turn, some organisms-specific FSFs such as the immunoglobulin superfamily that is specific to eukaryotes is not the most ancient fold despite its high abundance in eukaryotic proteomes. Thus, both FSF abundance and spread in modern proteomes determine the evolutionary age of each FSF on a relative timescale from 0 (most ancient) to 1 (most recent). It can be argued, however, that narrow spread of some FSFs could be due to evolutionary bottleneck or sampling biases. However, these events are difficult to compute and are minimized by focusing on the entire FSF repertoire. In other words, we believe that an artificial increase or decrease in the abundance and spread of some FSFs will not drastically affect the evolutionary picture as the phylogenetic model is driven from the global analysis.

The analysis revealed interesting trends in the evolution of cells and viruses that need to be individually described in three aspects:

FSF numbers suggest vertical traces in the evolution of cells and viruses. There are 15 possible Venn combinations for FSFs corresponding to the four supergroups of life, Archaea (A), Bacteria (B), Eukarya (E), and viruses (V) (Figure 6.2). The size of each Venn group is representative of the strength of evolutionary relationship between supergroups [256]. When coupled with evolutionary information, this exercise portrays key events in the evolutionary history of organisms and serves the purpose of a ToL, without tree reconstruction. The typical assumption is that shared features indicate common origin. For example, cells share a number of features that support their common ancestry. These include sharing a core of universally conserved genes, possessing lipid membranes and ribosomes, and the ability to carry out metabolism. Our data extend this idea to the molecular level and reveal that about one-fourth (492 out of 1,993) of the total FSFs were shared only by cellular organisms, ranging from microbial akaryotes to vertebrates and humans (ABE group). This is strong support for the common origin of cellular organisms. We argue that a similar logic could be extended to viruses as another one-fifth (395 out of 1,993) of the total FSFs were shared by cells and viruses of all replicon types (ABEV group). In fact, the ABEV group included both the very ancient and very recent FSFs (*nd* range from 0 to 1), however, it was mostly enriched with FSFs of ancient origin (median $nd = 0.35$). This suggests that three or four of the supergroups retained roughly 45% of the total FSFs. The most simple and parsimonious explanation for the very large sizes of the ABE and ABEV groups is the vertical evolution of cells and viruses from a cellular ancestor

(read below). Importantly, ABEV FSFs were detected in the proteomes of viruses harboring different replicon types (i.e. dsDNA, ssDNA, ssRNA, dsRNA, and retrotranscribing) as defined by the Baltimore classification scheme [213]. This indicates that all types of replication strategies were utilized in ancient cells suggesting that most of the modern DNA replication proteins and perhaps DNA itself was invented by ancient viruses (*sensu* [194,314]).

Evolution of viruses and Archaea by reductive loss of ancient FSFs. Most of the 15 Venn groups appeared at different times in evolution and in every instance represented key historical events. For example, the ABE group directly followed the ABEV group (Figure 6.2). The most ancient ABE FSF is the membrane transport protein 'MetI-like' FSF (f.58.1) that is a highly abundant cellular protein, especially in akaryotes. It was detected in 99% archaeal, 98% bacterial, and 9% eukaryal proteomes that were sampled. However, it was completely absent from the viral proteomes. We explain this as a result of reductive evolution. We argue that absence of an ancient (and highly abundant) FSF in only one out of the four supergroups is likely a 'loss' in one rather than 'gain' in three supergroups, as the latter scenario is comparatively less parsimonious. This is especially true if the particular FSF is widespread in the members of individual supergroups. As explained above, reductive evolution seems a more realistic scenario to explain viral evolution given their lifestyle resemblance with cellular parasites [219,220,267]. Reductive evolutionary tendencies were later experienced by archaeal organisms when the 'Lysozyme-like' FSF (d.2.1) was completely lost from Archaea at *nd* = 0.15 (BEV group). Again, d.2.1 was widespread in the remaining groups with 15%, 93%, and 73% presence in viral, bacterial, and eukaryal proteomes. Collectively, our data and timeline diagram highlight an early cellular existence of viruses and the onset of reductive evolutionary trends in the genomes and proteomes of the emerging members of the first supergroups, viruses and Archaea [20,86].

Appearance of novel FSFs and diversified modern cells and parasitic viruses. Venn groups and evolutionary timelines of protein domains also support the expected transformation of the protein world, from initial innovation tailored by reductive loss to unique repertoires confined to the cellular and viral members of the emerging supergroups (Figure 6.2). The new evolutionary phase involved the appearance of supergroup-specific FSFs that cannot evolve by HGT and uniquely identify groups of cellular organisms and viruses. All such gains occurred late in evolution, first in Bacteria (*nd* = 0.46), and then in viruses, Archaea, and Eukarya (*nd* = 0.59). Remarkably, the group of 67 virus-specific FSFs that lacked homologs in cellular proteomes (V

group) appeared together in a very small 'window' of the timeline (*nd* = 0.59–0.64). Their large number suggests that: (i) viruses are capable of creating genetic novelty [200,201], and (ii) viral proteomes do not solely grow by HGT. Moreover, viral-specific FSFs included most capsid/coat proteins and pathogenicity-related domains required to successfully infect cellular organisms (unpublished data).

Strikingly, the close appearances of AV, BV, and EV FSFs once repertoires specific to their respective cellular supergroups diversified in evolution strengthen the expected link between parasitic life cycles and emerging organismal lineages. The FSFs of these Venn groups were likely transferred from cells to modern viruses (or *vice versa*) via HGT and/or were structural innovations that helped establish viral infection cycles. Finally, because viral genomes can become part of cellular genomes, we expect that viral proteins are not restricted to the 'V' group. In turn, the AV, BV, EV, and ABEV groups likely include many proteins of viral origin that were taken from endogenized viruses or via HGT. Thus, viral proteins may be spread out in other groups and the actual number of proteins of viral origin may be even higher than the one reported in our study.

tRNA molecules reveal the early origin of RNA viruses

Only few RNA families are universal, and out of these, the tRNA family is assumed to be the most ancient [184]. Its structural makeup carries deep evolutionary history [106,347]. A phylogeny reconstructed from the sequence and structure of 571 tRNA molecules, however, failed to reveal a ToL with clear groupings of viral and cellular tRNAs [89] (Figure 6.3A). Instead, it placed molecules with a variable arm at the base of the tree. This probably stems from multiple episodes of structural and functional recruitment in the history of this molecule.

In order to uncover patterns of origin and evolution of lineages, tRNAs were forced into monophyletic groups (i.e. groups sharing a common ancestor) by restricting the search of optimal trees to pre-specified tree topologies [89]. The number of additional steps (S) required to constrain taxa into a variety of alternative groups (a selected set is shown in Figure 6.3B) define lineage coalescence distances and was used to build timelines (showing S increasing with elapsed time; open circles) or test alternative hypotheses of origin by selecting the most parsimonious (blue circles) (Figure 6.3C). For example, the tree of tRNAs (10,083 steps) was forced to fulfill the $((A,B,E),V_E,V_B)$ constraint of pooling tRNA from Archaea (A), Bacteria (B)

and Eukarya (E) into a single group and leaving viral tRNAs of eukaryotic (V_E) or bacterial (V_B) origin unconstrained (Figure 6.3A). Building this tree required many additional steps (S = 367). Figure 6.3C shows remarkable patterns obtained from this analysis. First, constraining each supergroup individually revealed the very early appearance of Archaea and viruses followed by the late appearance of Eukarya and Bacteria. These results match timelines obtained from phylogenomic analyses of FSFs (Figure 6.2) and considerable additional evidence [271]. Second, tRNA from eukaryoviruses that include several avian and murine RNA-based retroviral lineages appeared earlier than bacterioviruses with dsDNA replication strategies. This suggests that RNA viruses originated earlier than DNA viruses, which matches the suggested transition from RNA-based to DNA-based genomes in cellular evolution. Finally, the late appearance of Eukarya and Bacteria, which provide hosts to the viral groups we sampled, suggests that viruses established modern viral lifecycles when diversified lineages in these supergroups appeared (red branches in the reconstructed tree; Figure 6.3C). This is congruent with the relatively late onset of viral-specific FSFs in the evolutionary timeline of domains (Figure 6.2).

A data-driven model of viral origins

Our comparative genomic and phylogenomic exercise provides a historical account of the evolution of cells and viruses. This account unfolds ~3.8 billion years of planetary history. In light of our data, we propose that viruses evolved from ancient cells by genome reduction. These ancient cells could be referred to as 'proto-virocells' that hosted viral replicons but lacked the ability to produce modern day virions. Thus they could be contrasted from modern day virocells [241] that produce elaborate virions (built from 'jelly-roll' and other capsid proteins). It does not mean that no form of virion was produced in the proto-virocells. Perhaps, the ancient virions were vesicles that transported viral genomes in and out of the cells. This vesicle secretion phenomenon that is widespread in modern cells is regarded as a tool of intercellular communication [348] and a potential contributor to viral infection [349] (see [350,351] for other vesicle-related scenarios of viral origins). Another possibility is that perhaps protein folds not present in modern viruses were used to build primitive virions (see Figure 6.4 for a pictorial model of this hypothesis of origin).

In other words, ancient viruses were different from modern day viruses and existed in the form of ancient cells co-existing with the ancestors of Archaea, Bacteria, and Eukarya. While the

concept of ‘an ancient cell harboring virus replicon’ is difficult to view, it is very similar to modern cells harboring endogenized viruses. Thus, viruses always interacted with cells. In the very beginning they were one component (supported by the large size of ABEV). Then the cell and viral components disintegrated (mediated by reductive evolution/early appearance of ABE). And finally, the situation can be restored today when the viral component (re)-takes control of a modern cell or becomes part of its genome (helped by V repertoire of virus-specific FSFs). This scenario adequately explains both the origin of virus-specific FSFs and FSFs with cellular homologs and is logical since all known cellular parasites also evolve in a similar way, i.e. genome reduction and becoming dependent on their hosts.

A cellular origin of viruses also seems necessary since the ABEV group that includes 395 FSFs (~20% of the total FSFs that were sampled) shared between all types of viruses and cells was the most ancient Venn group (Figure 6.2). Interestingly, most of the ancient ABEV FSFs are members of proteins that are associated with membranes. Similarly, modern viruses are intimately associated with proteins (e.g. capsids) thus prerequisiting the existence of some sort of basic cell structure to support rudimentary metabolism and translation. A corollary is the existence of cells of different size at *nd* ~ 0.1 (~3.3 billions years ago). Remarkably, microfossil evidence in black chert beds and in shallow marine siliclastic deposits of that age revealed cellular microstructures of two broad size ranges, ~5-25 μm and ~300 μm in size [315,316,352]. We contend that microfossil size variation represent coexisting primordial cellular lines.

Taken together, these observations refute the idea of a ‘pre-cellular’ origin of viruses that is incompatible with virus biology (since viruses by definition are dependent upon cells for reproduction; for a new version see [207]). It is logical to think that many kinds of cells started in evolution but did not make it this far. The known survivors of billions of years of evolution are the three kinds of ribocells, Archaea, Bacteria, and Eukarya, and viruses that likely originated from a ‘fourth’ sibling of the ribocells. Having said this, viruses (or precisely the proto-virocells) started to lose FSFs very early in evolution. This is demonstrated by the appearance of the ABE group soon after ABEV. The ABE group includes FSFs found in all three kinds of ribocells but not in any modern viruses. We argue that the appearance of ABE is actually loss of V, suggesting that the probability of three independent gains is less likely than loss in one. By this argument, we propose the early scenario of reductive evolution in viruses. Reductive evolution is a near-universal phenomenon in all known cellular species that have become obligate parasites.

It is logical to think that viruses would evolve in a similar way, especially if they started to infect cells very early in evolution as shown by our data and as previously argued (see [219,220,267]).

One criticism to the reductive scenario of viral origins is that it cannot explain the origin of viral capsid proteins that are believed absent in cells. Several distinct capsid protein folds have now been characterized. One of such folds, the 'jelly-roll' fold is widespread in icosahedral viruses and especially abundant in the RNA viruses. However, the structural relatives of 'jelly-roll' are found in cells, especially the histone chaperones that assist in loading DNA onto histones [274,276]. Similarly, icosahedral structures that are morphologically similar to viral capsids have also been detected in akaryotic cells. Interestingly, these so-called protein compartments store enzymes instead of viral capsids that store nucleic acids. It has been hypothesized that perhaps a switch from storing enzymes to storing nucleic acids led to the origin of viral capsids in an ancient cell [287]. One example is the encapsulin protein that forms the protein shell of archaeal nanocompartments. Interestingly, encapsulin shares homologous domains with *Caudovirales* [287]. This clearly shows an overlap between both capsid folds and capsid-like structures in viruses and cells. Another protein compartment, the bacterial carboxysome, is also icosahedral and resembles viral capsids in morphology [288]. However, it is built from a fold not yet seen in any extant viruses. It is possible that a virus harboring this fold is yet to be discovered. Another likely scenario could be the loss of this fold from modern viruses and that would imply that it was an ancient capsid protein fold utilized by ancient viruses to infect ancient cells (thus supporting our hypothesis). Thus, we argue that capsid-like structures are not so alien to cells as generally thought. These recent findings suggest that our knowledge about the spread of viral capsids is very limited. Discovery of novel viruses from atypical habitats and many different hosts will definitely improve our knowledge about the virosphere. The last few years have seen a dramatic increase in the discovery of 'giant' viruses with genomes reaching up to 2.5Mb (e.g. pandoraviruses [215]). Recently, the genome of a mimivirus relative, the brown tide virus (AaV), was also reported (~380 Kb) [353]. About 47% of the 377 putative AaV proteins lacked any homologs to the NCBI *nr* database. For the rest 53%, authors established a cellular origin based on sequence similarity with known cellular and NCLDV proteins. Although the authors concluded that the ancestral virus had an even smaller genome and it grew by capturing genes from different sources, the analysis ignores the large amount of AaV genes without cellular homologs. Importantly, the proteome of AaV could be divided into

roughly two equal segments with one of unknown origin and the other exhibiting similarity to known proteins. The discovery of AaV confirms the existence of a continuum in the genome sizes of very small viruses and their giant outliers. It is expected that perhaps a near linear pattern in the genome sizes of viruses ranging from few Kbs to Mbs will be reached with the discovery of novel viruses.

We reconcile our hypothesis with the distribution of modern infectious viruses in host organisms and propose: (i) an early unfolding of the primordial virocell stem line harboring RNA genomes without engaging in parasitic lifestyles (since RNA virocells were likely more ancient than DNA virocells), (ii) the rise of archaeoviruses and bacterioviruses unfolding the DNA mode that was recruited into cells (the likely connect to a RNA-to-DNA transition in cells), and (iii) the rise of incredible RNA viral diversity in eukaryotes.

What lies ahead?

In order to reveal the true story behind viral origins and evolution we suggest focusing on three promising lines of research. These include learning how the different groups of viruses infect some hosts but not others, determining the structural and morphological similarities of viruses infecting hosts separated by large evolutionary distances, and studying molecular or organizational features that are highly conserved. It would be of importance to determine if viruses infecting distantly related hosts also share a significant number of FSFs, as it could provide strong support to the ancient origin of viruses. SCOP structural domains are conserved evolutionary units and therefore constitute powerful tools for retrodiction. However, there are other features in molecules and biological makeup that are also highly conserved. For example, our explorations could be supplemented by a similar analysis of evolution of molecular functions defined by the Gene Ontology database [58,59]. Perhaps the major problem in studying viral evolution is the widespread ‘belief’ that viruses are merely ‘gene robbers’ (as claimed in [237]). This is now challenged with comparative genomic data and the existence of virus-specific protein folds. However, it will be necessary to determine the proportion of vertically and horizontally inherited FSFs in each of the 15 Venn groups and other conserved features that may be useful for viral research. This will ensure the robustness of inferences drawn in the present review.

Figures

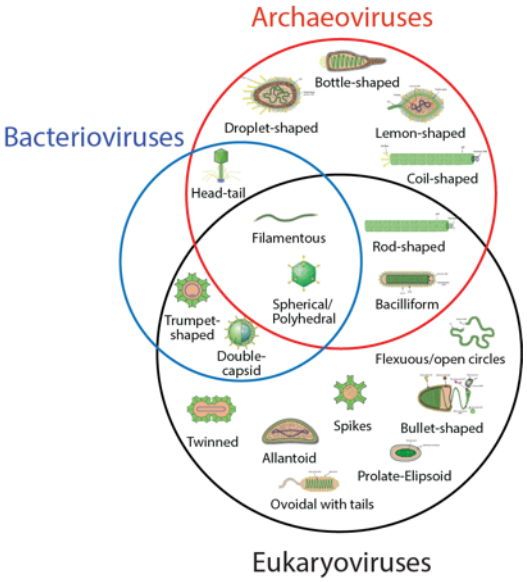

Figure 6.1 Virion morphotypes shared between and unique to archaeoviruses, bacterioviruses and eukaryoviruses. The Venn diagram shows that no morphotype was unique to bacterioviruses (modified from **Figure 5.1** in Chapter 5). Virion pictures were taken from ViralZone [320].

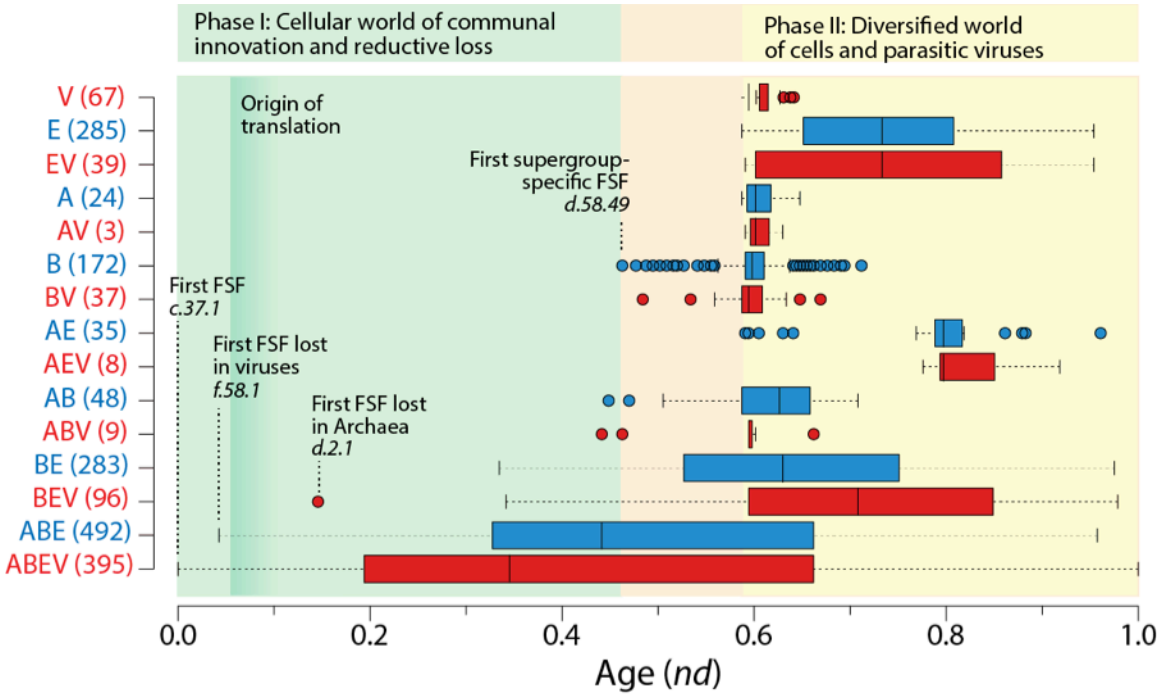

Figure 6.2 FSF domain sharing and evolution in cells and viruses. FSFs in each of 15 possible Venn distribution groups along a timeline of protein domain evolution, with domain age (*nd*) defined by the relative number of nodes in lineages of a tree of FSF domains. A total of 1,993 significant FSF domains (*E*-value < 0.0001) were detected when 4,211 completely sequenced viral and cellular proteomes were searched against SUPERFAMILY hidden Markov models of structure assignment [40,41]. The viral dataset included 1,125 dsDNA, 453 ssDNA, 122 dsRNA, 806 plus-ssRNA, 95 minus-ssRNA, and 114 retrotranscribing viruses. The cellular dataset included proteomes from 114 Archaea, 1,062 Bacteria, and 320 Eukarya. Viral and cellular groups were colored red and blue, respectively. The tree of domains used to construct the timeline was the single most parsimonious tree [tree length = 857,984, Consistency Index (CI) = 0.11, Retention Index (RI) = 0.78; Rescaled CI = 0.09; g_1 = -0.06] and described the evolution of 1,993 FSFs (taxa) using 4,211 proteomes (characters). The character states were normalized abundance values for each FSF in every proteome. Specifically, the raw abundance value of each FSF in every proteome was log-transformed and normalized by the maximum abundance value in the entire matrix using the following equation (g_{ab_norm} = Round [ln(g_{ab}+1) / ln(g_{max}+1) * 23]; see [29] for additional detail). The transformation takes care of the differences in genome sizes and unequal variances. The normalized values were rescaled from 0 to 23 to yield 24 possible character states that were compatible with PAUP (ver. 4.0b10) software [44]. The most parsimonious tree was calculated by a heuristic-based maximum parsimony search. We note that maximum parsimony performs better than likelihood methods when the analysis involves a large number of characters evolving at different rates [101]. The tree length is expected to be higher in an analysis of this magnitude involving 1,993 taxa and 4,211 characters. FSFs marking the onset of each major evolutionary event are labeled. FSF c.37.1 is the 'P-loop containing NTP hydrolase' FSF. Other FSFs described in text. Vertical bars within each distribution indicate group medians. FSF numbers are given in parenthesis. The two major evolutionary phases are highlighted in different background.

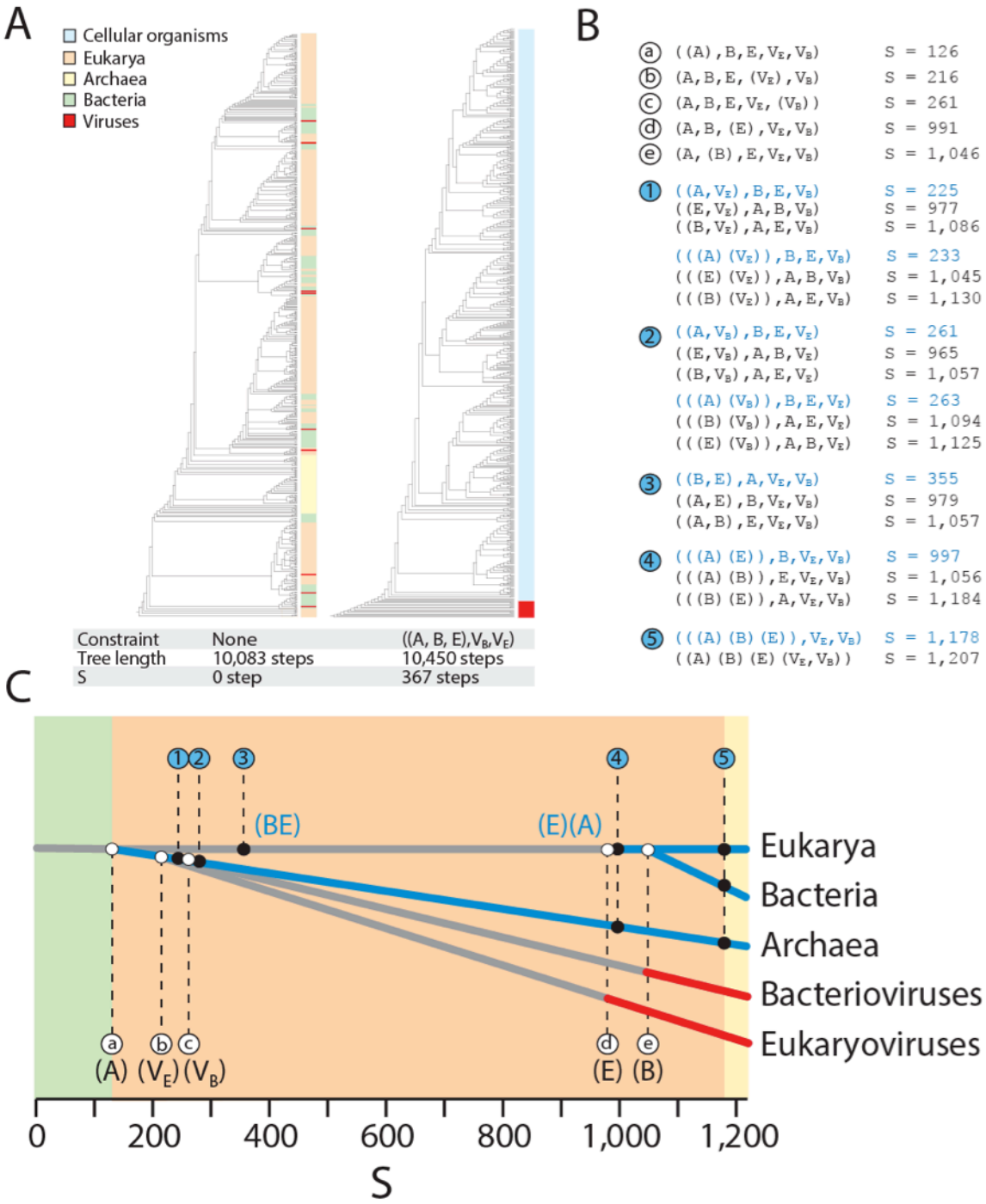

Figure 6.3 Evolutionary timelines of supergroups inferred from the sequence and structure of 571 tRNA molecules. Data were retrieved from the Bayreuth tRNA Database (http://www.staff.uni-bayreuth.de/~btc914/search/index.html, Part 2: compilation of tRNA sequences; September 2004 edition). Constraint analyses were conducted to search for the optimal trees based on pre-specified tree topologies using the "enforce topological constraint" option of PAUP* [44]. The number of additional steps (S) required to force (constrain) particular taxa into a monophyletic group were used to define an evolutionary distance with which to evaluate alternative phylogenetic hypotheses or to compare hypotheses that are not mutually exclusive. The latter approach was used to construct evolutionary timelines, in which lower S values corresponded to ancient tRNAs, a trend that was derived from the rooted trees (and embedded assumptions of polarization). Constraints were based on

grouping of tRNA molecules by organismal superkingdoms (A = Archaea, B = Bacteria, E = Eukarya) or viruses (V = viruses, V_B = viruses associated with Bacteria, V_E = viruses associated with Eukarya) using maximum parsimony analyses of combined tRNA structure and sequence data. Each constrained group is given in parentheses. The length of the most parsimonious tree derived from the combined data set was 10,083 steps [CI = 0.069 and 0.069, with and without uninformative characters, respectively; RI = 0.681; rescaled CI = 0.047; g_1 = 20.107]. **A)** Example of constraint analysis showing how forcing cellular supergroups into monophyly adds 367 additional steps to the most parsimonious tree reconstruction. **B)** Subset of constraint definitions and associated S values used in the analysis. **C)** Timeline of supergroup diversification showing how constraint representing non-competing (open circles) and competing (blue circles) hypotheses illustrate most parsimonious lineage relationships and their coalescence. Shaded areas of the timeline are delimited by lineage coalescence and describe three evolutionary epochs. The branch segment in red indicates the overlap of viral and diversified cellular history and suggests the late appearance of modern viral life cycles.

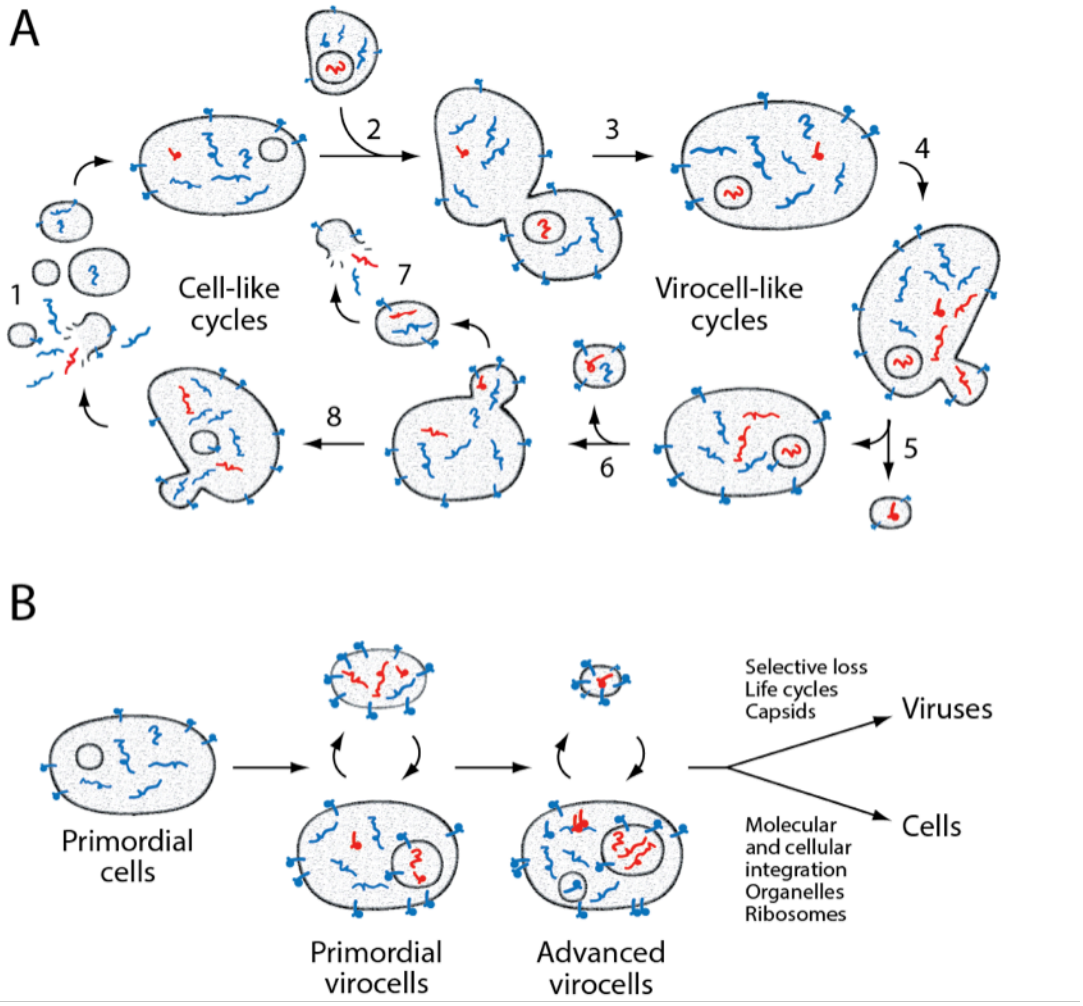

Figure 6.4 Model explaining the origin and early evolution of viruses. A) The illustration describes selected aspects of the complex dynamics of vesicle-entrapment of peptides and proteins (blue backbone traces), nucleic acids (red traces) and other vesicles in primordial cells. Vesicles behave as bioreactors hosting peptides and proteins, which sometimes insert into membranes and increase vesicle stability. Vesicles also regulate surface area through thermal energy, increasing their surface by accretion of amphiphilic hydrocarbons (feature 1) or by fusion with other vesicles (2 and 3), and reducing it by shedding microvesicles (4 and 5). Vesicles can also entrap smaller vesicles (similar to liposomes). These intracellular vesicles can fuse to other internal vesicles, increasing their surface and exchanging their contents. They can also be released from primordial cells by rupture of their hosts or by budding from membranes (6). If these microvesicles are not stable they will burst releasing their contents to the surrounding environment (7 and 8). We propose that microvesicles harboring nucleic acids (mostly primordial RNA genomes derived from tRNA) and stabilized by membrane proteins could have established virocell-like cycles (right) in which virion-like vesicles exchange genetic materials between the primordial cells. These cycles externalized the genome of the primordial virocells. Other cellular systems refrained from exchanging materials in this way, benefiting instead from encapsulation of nucleic acids released by bursting vesicles and their internalization. These cell-like cycles (left) preserved genetic materials inside primordial cells, in internal vesicles or in protoplasm. **B)** The cartoon describes the birth of modern cells and viruses from primordial replication strategies. Virocell-like replication favored selective loss of the molecular repertoires of the microvesicles, better stabilization via specialized capsid-like proteins, and elaborate life cycle dependencies. Molecular loss started early (ABE Venn group; *nd* ~ 0.04) but these reductive evolutionary tendencies continued throughout the timeline of domains. Cell-like replication focused instead on growth, diversification and cellular integration of internalized vesicles and their

genetic materials culminating in acidocalcisomes (present in all cellular supergroups), specialized organelles, and the nucleus. Cells also developed sophisticated molecular machinery, beginning with the catalytic domains of aminoacyl-tRNA synthetases of the translation apparatus (at $nd \sim 0.05$) and ending with a multifunctional ribosomal ensemble [29]. Thus, our hypothesis accounts for the development of capsids and ribosomes in two separate cellular stem lines responsible for modern virocells and ribocells, respectively [195].

BIBLIOGRAPHY

1. Toll-Riera M, Alba MM. (2013) Emergence of novel domains in proteins. BMC Evol Biol 13: 47-2148-13-47.
2. Moore AD, Bjorklund AK, Ekman D, Bornberg-Bauer E, Elofsson A. (2008) Arrangements in the modular evolution of proteins. Trends Biochem Sci 33: 444-451.
3. Chothia C, Gough J, Vogel C, Teichmann SA. (2003) Evolution of the protein repertoire. Science 300: 1701-1703.
4. Elofsson A, Sonnhammer EL. (1999) A comparison of sequence and structure protein domain families as a basis for structural genomics. Bioinformatics 15: 480-500.
5. Caetano-Anolles G, Wang M, Caetano-Anolles D, Mittenthal J. (2009) The origin, evolution and structure of the protein world. Biochem J 417: 621-637.
6. Wang M, Caetano-Anollés G. (2009) The evolutionary mechanics of domain organization in proteomes and the rise of modularity in the protein world. Structure 17: 66-78.
7. Moore AD, Bornberg-Bauer E. (2012) The dynamics and evolutionary potential of domain loss and emergence. Mol Biol Evol 29: 787-796.
8. Collins RE, Merz H, Higgs PG. (2011) Origin and evolution of gene families in bacteria and archaea. BMC Bioinformatics 12 Suppl 9: S14.
9. Hahn MW, Han MV, Han SG. (2007) Gene family evolution across 12 drosophila genomes. PLoS Genet 3: e197.
10. Koonin EV, Makarova KS, Aravind L. (2001) Horizontal gene transfer in prokaryotes: Quantification and classification. Annu Rev Microbiol 55: 709-742.
11. Buljan M, Bateman A. (2009) The evolution of protein domain families. Biochem Soc Trans 37: 751-755.
12. Ibba M, Curnow AW, Soll D. (1997) Aminoacyl-tRNA synthesis: Divergent routes to a common goal. Trends Biochem Sci 22: 39-42.
13. O'Donoghue P, Luthey-Schulten Z. (2003) On the evolution of structure in aminoacyl-tRNA synthetases. Microbiol Mol Biol Rev 67: 550-573.
14. Nasir A, Kim KM, Caetano-Anolles G. (2012) Giant viruses coexisted with the cellular ancestors and represent a distinct supergroup along with superkingdoms archaea, bacteria and eukarya. BMC Evol Biol 12: 156.
15. Kim HS, Mittenthal JE, Caetano-Anolles G. (2013) Widespread recruitment of ancient domain structures in modern enzymes during metabolic evolution. J Integr Bioinform 10: 214.
16. Kettler GC, Martiny AC, Huang K, Zucker J, Coleman ML, et al. (2007) Patterns and implications of gene gain and loss in the evolution of prochlorococcus. PLoS Genet 3: e231.
17. Zhu B, Zhou S, Lou M, Zhu J, Li B, et al. (2011) Characterization and inference of gene gain/loss along burkholderia evolutionary history. Evol Bioinform Online 7: 191-200.
18. Punta M, Coggill PC, Eberhardt RY, Mistry J, Tate J, et al. (2012) The pfam protein families database. Nucleic Acids Res 40: D290-301.
19. Pal LR, Guda C. (2006) Tracing the origin of functional and conserved domains in the human proteome: Implications for protein evolution at the modular level. BMC Evol Biol 6: 91.
20. Wang M, Yafremava LS, Caetano-Anollés D, Mittenthal JE, Caetano-Anollés G. (2007) Reductive evolution of architectural repertoires in proteomes and the birth of the tripartite world. Genome Res 17: 1572-1585.
21. Nasir A, Naeem A, Khan MJ, Nicora HDL, Caetano-Anollés G. (2011) Annotation of protein domains reveals remarkable conservation in the functional make up of proteomes across superkingdoms. Genes 2: 869-911.
22. Georgiades K, Merhej V, El Karkouri K, Raoult D, Pontarotti P. (2011) Gene gain and loss events in rickettsia and orientia species. Biol Direct 6: 6.
23. Zmasek CM, Godzik A. (2011) Strong functional patterns in the evolution of eukaryotic genomes revealed by the reconstruction of ancestral protein domain repertoires. Genome Biol 12: R4.
24. Hughes AL, Friedman R. (2004) Shedding genomic ballast: Extensive parallel loss of ancestral gene families in animals. J Mol Evol 59: 827-833.
25. Jain R, Rivera MC, Lake JA. (1999) Horizontal gene transfer among genomes: The complexity hypothesis. Proc Natl Acad Sci U S A 96: 3801-3806.
26. Treangen TJ, Rocha EP. (2011) Horizontal transfer, not duplication, drives the expansion of protein families in prokaryotes. PLoS Genet 7: e1001284.
27. Kim KM, Caetano-Anollés G. (2012) The evolutionary history of protein fold families and proteomes confirms that the archaeal ancestor is more ancient than the ancestors of other superkingdoms. BMC Evol Biol 12: 13.
28. Caetano-Anollés G, Caetano-Anollés D. (2003) An evolutionarily structured universe of protein architecture. Genome Res 13: 1563-1571.

29. Caetano-Anollés D, Kim KM, Mittenthal JE, Caetano-Anollés G. (2011) Proteome evolution and the metabolic origins of translation and cellular life. J Mol Evol 72: 14-33.
30. Yang S, Doolittle RF, Bourne PE. (2005) Phylogeny determined by protein domain content. Proc Natl Acad Sci U S A 102: 373-378.
31. Lin J, Gerstein M. (2000) Whole-genome trees based on the occurrence of folds and orthologs: Implications for comparing genomes on different levels. Genome Res 10: 808-818.
32. Zhang Y, Chandonia JM, Ding C, Holbrook SR. (2005) Comparative mapping of sequence-based and structure-based protein domains. BMC Bioinformatics 6: 77.
33. Andreeva A, Howorth D, Chandonia JM, Brenner SE, Hubbard TJ, et al. (2008) Data growth and its impact on the SCOP database: New developments. Nucleic Acids Res 36: D419-25.
34. Murzin AG, Brenner SE, Hubbard T, Chothia C. (1995) SCOP: A structural classification of proteins database for the investigation of sequences and structures. J Mol Biol 247: 536-540.
35. Illergård K, Ardell DH, Elofsson A. (2009) Structure is three to ten times more conserved than sequence—a study of structural response in protein cores. Proteins 77: 499-508.
36. Müller A, MacCallum R, Sternberg M. (2002) Structural characterization of the human proteome. Genome Res 12: 1625-1641.
37. Caetano-Anollés G, Nasir A. (2012) Benefits of using molecular structure and abundance in phylogenomic analysis. Front Genet 3: 172.
38. Kim KM, Caetano-Anollés G. (2011) The proteomic complexity and rise of the primordial ancestor of diversified life. BMC Evol Biol 11: 140.
39. Wilson D, Pethica R, Zhou Y, Talbot C, Vogel C, et al. (2009) SUPERFAMILY--sophisticated comparative genomics, data mining, visualization and phylogeny. Nucleic Acids Res 37: D380-6.
40. Gough J, Chothia C. (2002) SUPERFAMILY: HMMs representing all proteins of known structure. SCOP sequence searches, alignments and genome assignments. Nucleic Acids Res 30: 268-272.
41. Gough J, Karplus K, Hughey R, Chothia C. (2001) Assignment of homology to genome sequences using a library of hidden markov models that represent all proteins of known structure. J Mol Biol 313: 903-919.
42. Wilson D, Madera M, Vogel C, Chothia C, Gough J. (2007) The SUPERFAMILY database in 2007: Families and functions. Nucleic Acids Res 35: D308-13.
43. Wang M, Caetano-Anollés G. (2006) Global phylogeny determined by the combination of protein domains in proteomes. Mol Biol Evol 23: 2444-2454.
44. Swofford DL. (2002) PAUP*: Phylogenetic analysis using parsimony (*and other methods), version 4.0b10. Sunderland, MA: Sinauer Associates.
45. Weston PH. (1988) Indirect and direct methods in systematics. In: Humphries CJ, ed. Ontogeny and Systematics. New York: Columbia University Press. pp. 27-56.
46. Weston PH. (1994) Methods for rooting cladistic trees. In: Siebert DJ, Scotland RW, Williams DM, editors. Models in Phylogeny Reconstruction. Oxford: Oxford University Press. pp. 125-155.
47. Lundberg JG. (1972) Wagner networks and ancestors. Syst Biol 21: 398-413.
48. Kitching IJ, Forey PL, Humphries CJ, Williams DM. (1998) Cladistics: The theory and practice of parsimony analysis. Oxford: Oxford University Press. 228 p.
49. Huson DH, Richter DC, Rausch C, Dezulian T, Franz M, et al. (2007) Dendroscope: An interactive viewer for large phylogenetic trees. BMC Bioinformatics 8: 460.
50. Puigbo P, Garcia-Vallve S, McInerney JO. (2007) TOPD/FMTS: A new software to compare phylogenetic trees. Bioinformatics 23: 1556-1558.
51. Wang M, Jiang Y, Kim KM, Qu G, Ji H, et al. (2011) A universal molecular clock of protein folds and its power in tracing the early history of aerobic metabolism and planet oxygenation. Mol Biol Evol 28: 567-582.
52. Caetano-Anollés K, Caetano-Anollés G. (2013) Structural phylogenomics reveals gradual evolutionary replacement of abiotic chemistries by protein enzymes in purine metabolism. PloS One 8: e59300.
53. Vogel C, Berzuini C, Bashton M, Gough J, Teichmann SA. (2004) Supra-domains: Evolutionary units larger than single protein domains. J Mol Biol 336: 809-823.
54. Vogel C, Teichmann SA, Pereira-Leal J. (2005) The relationship between domain duplication and recombination. J Mol Biol 346: 355-365.
55. Vogel C, Chothia C. (2006) Protein family expansions and biological complexity. PLoS Comput Biol 2: e48.
56. Fang H, Gough J. (2013) DcGO: Database of domain-centric ontologies on functions, phenotypes, diseases and more. Nucleic Acids Res 41: D536-44.
57. de Lima Morais DA, Fang H, Rackham OJ, Wilson D, Pethica R, et al. (2011) SUPERFAMILY 1.75 including a domain-centric gene ontology method. Nucleic Acids Res 39: D427-34.

58. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, et al. (2000) Gene ontology: Tool for the unification of biology. Nat Genet 25: 25-29.
59. Harris M, Clark J, Ireland A, Lomax J, Ashburner M, et al. (2004) The gene ontology (GO) database and informatics resource. Nucleic Acids Res 32: D258-61.
60. Benjamini YH, Y. (1995) Controlling the false discovery rate - a practical and powerful approach to multiple testing. J R Stat Soc 57: 289-300.
61. Jones PM, George AM. (2004) The ABC transporter structure and mechanism: Perspectives on recent research. Cell Mol Life Sci 61: 682-699.
62. Davidson AL, Dassa E, Orelle C, Chen J. (2008) Structure, function, and evolution of bacterial ATP-binding cassette systems. Microbiol Mol Biol Rev 72: 317-64, table of contents.
63. Large AT, Goldberg MD, Lund PA. (2009) Chaperones and protein folding in the archaea. Biochem Soc Trans 37: 46-51.
64. Yafremava LS, Wielgos M, Thomas S, Nasir A, Wang M, et al. (2013) A general framework of persistence strategies for biological systems helps explain domains of life. Front Genet 4: 16.
65. Koonin EV. (2010) The origin and early evolution of eukaryotes in the light of phylogenomics. Genome Biol 11: 209.
66. López-García P, Moreira D. (1999) Metabolic symbiosis at the origin of eukaryotes. Trends Biochem Sci 24: 88-93.
67. Martin W, Müller M. (1998) The hydrogen hypothesis for the first eukaryote. Nature 392: 37-41.
68. Gray MW. (2012) Mitochondrial evolution. Cold Spring Harb Perspect Biol 4. a011403.
69. Rivera MC, Lake JA. (2004) The ring of life provides evidence for a genome fusion origin of eukaryotes. Nature 431: 152-155.
70. Forterre P. (2013) The common ancestor of archaea and eukarya was not an archaeon. Archaea 2013: 372396.
71. Kelman Z. (2000) DNA replication in the third domain (of life). Curr Protein Pept Sci 1: 139-154.
72. Grabowski B, Kelman Z. (2003) Archeal DNA replication: Eukaryal proteins in a bacterial context. Annu Rev Microbiol 57: 487-516.
73. Sandman K, Reeve JN. (2000) Structure and functional relationships of archaeal and eukaryal histones and nucleosomes. Arch Microbiol 173: 165-169.
74. Woese CR. (1987) Bacterial evolution. Microbiol Rev 51: 221-271.
75. Bukhari SA, Caetano-Anollés G. (2013) Origin and evolution of protein fold designs inferred from phylogenomic analysis of CATH domain structures in proteomes. PLoS Comput Biol 9: e1003009.
76. Margulis L, Chapman M, Guerrero R, Hall J. (2006) The last eukaryotic common ancestor (LECA): Acquisition of cytoskeletal motility from aerotolerant spirochetes in the proterozoic eon. Proc Natl Acad Sci U S A 103: 13080-13085.
77. Cavalier-Smith T. (2002) The phagotrophic origin of eukaryotes and phylogenetic classification of protozoa. Int J Syst Evol Microbiol 52: 297-354.
78. Kurland C, Collins L, Penny D. (2006) Genomics and the irreducible nature of eukaryote cells. Science 312: 1011-1014.
79. de Duve C. (2007) The origin of eukaryotes: A reappraisal. Nat Rev Genet 8: 395-403.
80. Woese C. (1998) The universal ancestor. Proc Natl Acad Sci U S A 95: 6854-6859.
81. Iwabe N, Kuma K, Hasegawa M, Osawa S, Miyata T. (1989) Evolutionary relationship of archaebacteria, eubacteria, and eukaryotes inferred from phylogenetic trees of duplicated genes. Proc Natl Acad Sci U S A 86: 9355-9359.
82. Gogarten JP, Taiz L. (1992) Evolution of proton pumping ATPases: Rooting the tree of life. Photosynthesis Res 33: 137-146.
83. Xue H, Ng S, Tong K, Wong J. (2005) Congruence of evidence for a methanopyrus-proximal root of life based on transfer RNA and aminoacyl-tRNA synthetase genes. Gene 360: 120-130.
84. Xue H, Tong K, Marck C, Grosjean H, Wong J. (2003) Transfer RNA paralogs: Evidence for genetic code-amino acid biosynthesis coevolution and an archaeal root of life. Gene 310: 59-66.
85. Di Giulio M. (2007) The tree of life might be rooted in the branch leading to nanoarchaeota. Gene 401: 108-113.
86. Wang M, Kurland CG, Caetano-Anollés G. (2011) Reductive evolution of proteomes and protein structures. Proc Natl Acad Sci U S A 108: 11954-11958.
87. Zmasek C, Godzik A. (2010) Evolution of the protein domain repertoire of eukaryotes reveals strong functional patterns. Genome Biol 11: 43.
88. Kim KM, Caetano-Anollés G. (2010) Emergence and evolution of modern molecular functions inferred from phylogenomic analysis of ontological data. Mol Biol Evol 27: 1710-1733.

89. Sun F, Caetano-Anollés G. (2008) Evolutionary patterns in the sequence and structure of transfer RNA: Early origins of archaea and viruses. PLoS Comput Biol 4: e1000018.
90. Sober E, Steel M. (2002) Testing the hypothesis of common ancestry. J Theor Biol 218: 395-408.
91. Morrison DA. (2009) Why would phylogeneticists ignore computerized sequence alignment? Syst Biol 58: 150-158.
92. Maddison WP. (1993) Missing data versus missing characters in phylogenetic analysis. Syst Biol 42: 576.
93. De Laet J. (2005) Parsimony and the problem of inapplicables in sequence data. In: Albert VA, editor. Parsimony, phylogeny and genomics. Oxford: Oxford University Press. pp. 81-116.
94. Kluge AG, Farris JS. (1969) Quantitative phyletics and the evolution of anurans. Syst Zool 18: 1-32.
95. Huelsenbeck JP, Nielsen R. (1999) Effect of nonindependent substitution on phylogenetic accuracy. Syst Biol 48: 317-328.
96. Anisimova M, Cannarozzi GM, Liberles DA. (2010) Finding the balance between the mathematical and biological optima in multiple sequence alignment. Trends Evol Biol 2: e7.
97. Harish A, Caetano-Anollés G. (2012) Ribosomal history reveals origins of modern protein synthesis. PLoS One 7: e32776.
98. Martin W, Embley TM. (2004) Evolutionary biology: Early evolution comes full circle. Nature 431: 134-137.
99. Thrash JC, Boyd A, Huggett MJ, Grote J, Carini P, et al. (2011) Phylogenomic evidence for a common ancestor of mitochondria and the SAR11 clade. Sci Rep 1: 13.
100. Poole AM, Neumann N. (2011) Reconciling an archaeal origin of eukaryotes with engulfment: A biologically plausible update of the eocyte hypothesis. Res Microbiol 162: 71-76.
101. Kolaczkowski B, Thornton JW. (2004) Performance of maximum parsimony and likelihood phylogenetics when evolution is heterogeneous. Nature 431: 980-984.
102. Delsuc F, Brinkmann H, Philippe H. (2005) Phylogenomics and the reconstruction of the tree of life. Nat Rev Genet 6: 361-375.
103. Woese CR, Maniloff J, Zablen LB. (1980) Phylogenetic analysis of the mycoplasmas. Proc Natl Acad Sci U S A 77: 494-498.
104. Sankoff D, Leduc G, Antoine N, Paquin B, Lang BF, et al. (1992) Gene order comparisons for phylogenetic inference: Evolution of the mitochondrial genome. Proc Natl Acad Sci U S A 89: 6575-6579.
105. Gu X, Zhang H. (2004) Genome phylogenetic analysis based on extended gene contents. Mol Biol Evol 21: 1401-1408.
106. Sun F, Caetano-Anollés G. (2008) The origin and evolution of tRNA inferred from phylogenetic analysis of structure. J Mol Evol 66: 21-35.
107. Woese CR, Fox GE. (1977) Phylogenetic structure of the prokaryotic domain: The primary kingdoms. Proc Natl Acad Sci U S A 74: 5088-5090.
108. Penny D, Poole A. (1999) The nature of the last universal common ancestor. Curr Opin Genet Dev 9: 672-677.
109. Koonin EV. (2003) Comparative genomics, minimal gene-sets and the last universal common ancestor. Nat Rev Microbiol 1: 127-136.
110. Forterre P, Philippe H. (1999) Where is the root of the universal tree of life? Bioessays 21: 871-879.
111. Pace NR. (2009) Mapping the tree of life: Progress and prospects. Microbiol Mol Biol Rev 73: 565-576.
112. Gerstein M. (1998) Patterns of protein-fold usage in eight microbial genomes: A comprehensive structural census. Proteins 33: 518-534.
113. Marcet-Houben M, Puigbo P, Romeu A, Garcia-Vallve S. (2007) Towards reconstructing a metabolic tree of life. Bioinformation 2: 135-144.
114. Chang CW, Lyu PC, Arita M. (2011) Reconstructing phylogeny from metabolic substrate-product relationships. BMC Bioinformatics 12 Suppl 1: S27.
115. Tatusov RL, Fedorova ND, Jackson JD, Jacobs AR, Kiryutin B, et al. (2003) The COG database: An updated version includes eukaryotes. BMC Bioinformatics 4: 41.
116. Overbeek R, Begley T, Butler RM, Choudhuri JV, Chuang HY, et al. (2005) The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes. Nucleic Acids Res 33: 5691-5702.
117. Jensen RA. (1976) Enzyme recruitment in evolution of new function. Annu Rev Microbiol 30: 409-425.
118. Khersonsky O, Tawfik DS. (2010) Enzyme promiscuity: A mechanistic and evolutionary perspective. Annu Rev Biochem 79: 471-505.
119. Rhee SY, Wood V, Dolinski K, Draghici S. (2008) Use and misuse of the gene ontology annotations. Nat Rev Genet 9: 509-515.
120. Warnow T. (2012) Standard maximum likelihood analyses of alignments with gaps can be statistically inconsistent. PLoS Curr 4: RRN1308.

121. Liolios K, Chen IA, Mavromatis K, Tavernarakis N, Hugenholtz P, et al. (2010) The genomes on line database (GOLD) in 2009: Status of genomic and metagenomic projects and their associated metadata. Nucleic Acids Res 38: D346-D354.
122. Gough J. (2005) Convergent evolution of domain architectures (is rare). Bioinformatics 21: 1464-1471.
123. Farris JS. (1989) The retention index and homoplasy excess. Syst Zool 38: 406.
124. Cummings MP, Neel MC, Shaw KL. (2008) A genealogical approach to quantifying lineage divergence. Evolution 62: 2411-2422.
125. Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, et al. (2013) The SILVA ribosomal RNA gene database project: Improved data processing and web-based tools. Nucleic Acids Res 41: D590-6.
126. Posada D, Crandall KA. (1998) MODELTEST: Testing the model of DNA substitution. Bioinformatics 14: 817-818.
127. Perelman P, Johnson WE, Roos C, Seuanez HN, Horvath JE, et al. (2011) A molecular phylogeny of living primates. PLoS Genet 7: e1001342.
128. Saitou N, Nei M. (1987) The neighbor-joining method: A new method for reconstructing phylogenetic trees. Mol Biol Evol 4: 406-425.
129. Bryant D, Moulton V. (2004) Neighbor-net: An agglomerative method for the construction of phylogenetic networks. Mol Biol Evol 21: 255-265.
130. Huson DH. (1998) SplitsTree: Analyzing and visualizing evolutionary data. Bioinformatics 14: 68-73.
131. Holland BR, Huber KT, Dress A, Moulton V. (2002) Delta plots: A tool for analyzing phylogenetic distance data. Mol Biol Evol 19: 2051-2059.
132. Wichmann K, Holman EW, Rama T, Walker RS. Correlates of reticulation in linguistic phylogenies. LDC 1: 205-240.
133. Buckley CD. (2012) Investigating cultural evolution using phylogenetic analysis: The origins and descent of the southeast asian tradition of warp ikat weaving. PLoS One 7: e52064.
134. Garcia-Vallve S, Guzmán E, Montero M, Romeu A. (2003) HGT-DB: A database of putative horizontally transferred genes in prokaryotic complete genomes. Nucleic Acids Res 31: 187-189.
135. Forslund K, Henricson A, Hollich V, Sonnhammer EL. (2008) Domain tree-based analysis of protein architecture evolution. Mol Biol Evol 25: 254-264.
136. Dufresne A, Garczarek L, Partensky F. (2005) Accelerated evolution associated with genome reduction in a free-living prokaryote. Genome Biol 6: R14.
137. Giovannoni SJ, Tripp HJ, Givan S, Podar M, Vergin KL, et al. (2005) Genome streamlining in a cosmopolitan oceanic bacterium. Science 309: 1242-1245.
138. Ravin NV, Mardanov AV, Beletsky AV, Kublanov IV, Kolganova TV, et al. (2009) Complete genome sequence of the anaerobic, protein-degrading hyperthermophilic crenarchaeon desulfurococcus kamchatkensis. J Bacteriol 191: 2371-2379.
139. Anderson I, Rodriguez J, Susanti D, Porat I, Reich C, et al. (2008) Genome sequence of thermofilum pendens reveals an exceptional loss of biosynthetic pathways without genome reduction. J Bacteriol 190: 2957-2965.
140. Zillig W, Holz I, Janekovic D, Klenk HP, Imsel E, et al. (1990) Hyperthermus butylicus, a hyperthermophilic sulfur-reducing archaebacterium that ferments peptides. J Bacteriol 172: 3959-3965.
141. Anderson IJ, Dharmarajan L, Rodriguez J, Hooper S, Porat I, et al. (2009) The complete genome sequence of staphylothermus marinus reveals differences in sulfur metabolism among heterotrophic crenarchaeota. BMC Genomics 10: 145.
142. Kloesges T, Popa O, Martin W, Dagan T. (2011) Networks of gene sharing among 329 proteobacterial genomes reveal differences in lateral gene transfer frequency at different phylogenetic depths. Mol Biol Evol 28: 1057-1074.
143. Dopazo H, Santoyo J, Dopazo J. (2004) Phylogenomics and the number of characters required for obtaining an accurate phylogeny of eukaryote model species. Bioinformatics 20 Suppl 1: i116-21.
144. Yang S, Bourne PE. (2009) The evolutionary history of protein domains viewed by species phylogeny. PLoS One 4: e8378.
145. Doolittle WF. (1999) Phylogenetic classification and the universal tree. Science 284: 2124-2128.
146. Vesteg M, Krajcovic J. (2008) Origin of eukaryotic cells as a symbiosis of parasitic alpha-proteobacteria in the periplasm of two-membrane-bounded sexual pre-karyotes. Commun Integr Biol 1: 104-113.
147. Caro-Quintero A, Deng J, Auchtung J, Brettar I, Hofle MG, et al. (2011) Unprecedented levels of horizontal gene transfer among spatially co-occurring shewanella bacteria from the baltic sea. ISME J 5: 131-140.
148. McDaniel LD, Young E, Delaney J, Ruhnau F, Ritchie KB, et al. (2010) High frequency of horizontal gene transfer in the oceans. Science 330: 50.

149. Aminov RI. (2011) Horizontal gene exchange in environmental microbiota. Front Microbiol 2: 158.
150. Orengo CA, Michie A, Jones S, Jones DT, Swindells M, et al. (1997) CATH–a hierarchic classification of protein domain structures. Structure 5: 1093-1109.
151. Sun F, Caetano-Anollés G. (2009) The evolutionary history of the structure of 5S ribosomal RNA. J Mol Evol 69: 430-443.
152. Sun F, Caetano-Anollés G. (2010) The ancient history of the structure of ribonuclease P and the early origins of archaea. BMC Bioinformatics 11: 153.
153. Nasir A, Kim KM, Caetano-Anollés G. A phylogenomic census of molecular functions identifies modern thermophilic archaea as the most ancient form of cellular life. Archaea 2014: 706468.
154. Wong J, Chen J, Mat W, Ng S, Xue H. (2007) Polyphasic evidence delineating the root of life and roots of biological domains. Gene 403: 39-52.
155. Gogarten JP, Olendzenski L. (1999) Orthologs, paralogs and genome comparisons. Curr Opin Genet Dev 9: 630-636.
156. Dagan T, Roettger M, Bryant D, Martin W. (2010) Genome networks root the tree of life between prokaryotic domains. Genome Biol Evol 2: 379.
157. Cavalier-Smith T. (2002) The neomuran origin of archaebacteria, the negibacterial root of the universal tree and bacterial megaclassification. Int J Syst Evol Microbiol 52: 7-76.
158. Bapteste E, Brochier C. (2004) On the conceptual difficulties in rooting the tree of life. Trends Microbiol 12: 9-13.
159. Lake JA, Skophammer RG, Herbold CW, Servin JA. (2009) Genome beginnings: Rooting the tree of life. Philos Trans R Soc Lond B Biol Sci 364: 2177-2185.
160. Szathmary E, Smith JM. (1995) The major evolutionary transitions. Nature 374: 227-232.
161. Jablonka E, Lamb MJ. (2006) The evolution of information in the major transitions. J Theor Biol 239: 236-246.
162. Gribaldo S, Brochier-Armanet C. (2006) The origin and evolution of archaea: A state of the art. Philos Trans R Soc Lond B Biol Sci 361: 1007-1022.
163. Brochier-Armanet C, Forterre P, Gribaldo S. (2011) Phylogeny and evolution of the archaea: One hundred genomes later. Curr Opin Microbiol 14: 274-281.
164. Brochier C, Philippe H. (2002) Phylogeny: A non-hyperthermophilic ancestor for bacteria. Nature 417: 244.
165. Woese CR, Kandler O, Wheelis ML. (1990) Towards a natural system of organisms: Proposal for the domains archaea, bacteria, and eucarya. Proc Natl Acad Sci U S A 87: 4576-4579.
166. Rappe MS, Giovannoni SJ. (2003) The uncultured microbial majority. Annu Rev Microbiol 57: 369-394.
167. Forterre P, Bouthier De La Tour C, Philippe H, Duguet M. (2000) Reverse gyrase from hyperthermophiles: Probable transfer of a thermoadaptation trait from archaea to bacteria. Trends Genet 16: 152-154.
168. Confalonieri F, Elie C, Nadal M, de La Tour C, Forterre P, et al. (1993) Reverse gyrase: A helicase-like domain and a type I topoisomerase in the same polypeptide. Proc Natl Acad Sci U S A 90: 4753-4757.
169. Forterre P, Bergerat A, Lopez-Garcia P. (1996) The unique DNA topology and DNA topoisomerases of hyperthermophilic archaea. FEMS Microbiol Rev 18: 237-248.
170. Gupta R. (2000) The phylogeny of proteobacteria: Relationships to other eubacterial phyla and eukaryotes. FEMS Microbiol Rev 24: 367-402.
171. Griffiths E, Gupta RS. (2004) Signature sequences in diverse proteins provide evidence for the late divergence of the order aquificales. Int Microbiol 7: 41-52.
172. Ciccarelli FD, Doerks T, Von Mering C, Creevey CJ, Snel B, et al. (2006) Toward automatic reconstruction of a highly resolved tree of life. Science 311: 1283-1287.
173. Loytynoja A, Milinkovitch MC. (2001) Molecular phylogenetic analyses of the mitochondrial ADP-ATP carriers: The plantae/fungi/metazoa trichotomy revisited. Proc Natl Acad Sci U S A 98: 10202-10207.
174. Emes RD, Goodstadt L, Winter EE, Ponting CP. (2003) Comparison of the genomes of human and mouse lays the foundation of genome zoology. Hum Mol Genet 12: 701-709.
175. Liu L, Pearl DK, Brumfield RT, Edwards SV. (2008) Estimating species trees using multiple-allele DNA sequence data. Evolution 62: 2080-2091.
176. Chappe B, Michaelis W, Albrecht P, Ourisson G. (1979) Fossil evidence for a novel series of archaebacterial lipids. Naturwissenschaften 66: 522-523.
177. Michaelis W, Albrecht P. (1979) Molecular fossils of archaebacteria in kerogen. Naturwissenschaften 66: 420-421.
178. Schopf JW. (1999) Deep divisions in the tree of life--what does the fossil record reveal? Biol Bull 196: 351-3; discussion 354-5.

179. Blank CE. (2009) Not so old archaea - the antiquity of biogeochemical processes in the archaeal domain of life. Geobiology 7: 495-514.
180. Gogarten JP, Kibak H, Dittrich P, Taiz L, Bowman EJ, et al. (1989) Evolution of the vacuolar H+-ATPase: Implications for the origin of eukaryotes. Proc Natl Acad Sci U S A 86: 6661-6665.
181. Gribaldo S, Philippe H. (2002) Ancient phylogenetic relationships. Theor Popul Biol 61: 391-408.
182. Philippe H, Forterre P. (1999) The rooting of the universal tree of life is not reliable. J Mol Evol 49: 509-523.
183. Popa O, Dagan T. (2011) Trends and barriers to lateral gene transfer in prokaryotes. Curr Opin Microbiol 14: 615-623.
184. Hoeppner MP, Gardner PP, Poole AM. (2012) Comparative analysis of RNA families reveals distinct repertoires for each domain of life. PLoS Comput Biol 8: e1002752.
185. Olsen GJ, Woese CR, Overbeek R. (1994) The winds of (evolutionary) change: Breathing new life into microbiology. J Bacteriol 176: 1.
186. Brown JR. (2003) Ancient horizontal gene transfer. Nat Rev Genet 4: 121-132.
187. Moissl-Eichinger C, Huber H. (2011) Archaeal symbionts and parasites. Curr Opin Microbiol 14: 364-370.
188. Ungar D, Hughson FM. (2003) SNARE protein structure and function. Annu Rev Cell Dev Biol 19: 493-517.
189. Gribaldo S, Poole AM, Daubin V, Forterre P, Brochier-Armanet C. (2010) The origin of eukaryotes and their relationship with the archaea: Are we at a phylogenomic impasse? Nat Rev Microbiol 8: 743-752.
190. Kuriyan J, O'Donnell M. (1993) Sliding clamps of DNA polymerases. J Mol Biol 234: 915-925.
191. Stillman B. (1994) Smart machines at the DNA replication fork. Cell 78: 725-728.
192. Kleman-Leyer K, Armbruster DW, Daniels CJ. (1997) Properties of H. volcanii tRNA intron endonuclease reveal a relationship between the archaeal and eucaryal tRNA intron processing systems. Cell 89: 839-847.
193. Koonin EV, Senkevich TG, Dolja VV. (2009) Compelling reasons why viruses are relevant for the origin of cells. Nat Rev Microbiol 7: 615; author reply 615.
194. Forterre P. (2006) The origin of viruses and their possible roles in major evolutionary transitions. Virus Res 117: 5-16.
195. Nasir A, Kim KM, Caetano-Anollés G. (2012) Viral evolution: Primordial cellular origins and late adaptation to parasitism. Mob Genet Elements 2: 247-252.
196. Desmond E, Brochier-Armanet C, Forterre P, Gribaldo S. (2011) On the last common ancestor and early evolution of eukaryotes: Reconstructing the history of mitochondrial ribosomes. Res Microbiol 162: 53-70.
197. Kandler O. (1993) Cell wall biochemistry and three-domain concept of life. Syst Appl Microbiol 16: 501-509.
198. Woese CR. (2002) On the evolution of cells. Proc Natl Acad Sci U S A 99: 8742-8747.
199. Raoult D, Forterre P. (2008) Redefining viruses: Lessons from mimivirus. Nat Rev Microbiol 6: 315-319.
200. Forterre P, Prangishvili D. (2009) The great billion-year war between ribosome-and capsid-encoding organisms (cells and viruses) as the major source of evolutionary novelties. Ann N Y Acad Sci 1178: 65-77.
201. Abroi A, Gough J. (2011) Are viruses a source of new protein folds for organisms? - virosphere structure space and evolution. Bioessays 33: 626-635.
202. Forterre P. (2012) Darwin's goldmine is still open: Variation and selection run the world. Front Cell Infect Microbiol 2: 106.
203. Forterre P. (2005) The two ages of the RNA world, and the transition to the DNA world: A story of viruses and cells. Biochimie 87: 793-803.
204. Takemura M. (2001) Poxviruses and the origin of the eukaryotic nucleus. J Mol Evol 52: 419-425.
205. Bell PJL. (2001) Viral eukaryogenesis: Was the ancestor of the nucleus a complex DNA virus? J Mol Evol 53: 251-256.
206. Lodish H, Berk A, Zipursky SL, Matsudaira P, Baltimore D, et al. (2000) Molecular Cell Biology, 4th edition. New York: W. H. Freeman and Company. 1184 p.
207. Koonin EV, Senkevich TG, Dolja VV. (2006) The ancient virus world and evolution of cells. Biol Direct 1: 29.
208. Suttle CA. (2007) Marine viruses--major players in the global ecosystem. Nat Rev Microbiol 5: 801-812.
209. Rohwer F, Thurber RV. (2009) Viruses manipulate the marine environment. Nature 459: 207-212.
210. Pietilä MK, Demina TA, Atanasova NS, Oksanen HM, Bamford DH. (2014) Archaeal viruses and bacteriophages: Comparisons and contrasts. Trends Microbiol 22: 334-344.
211. Pina M, Bize A, Forterre P, Prangishvili D. (2011) The archeoviruses. FEMS Microbiol Rev 35: 1035-1054.
212. Nasir A, Forterre P, Kim KM, Caetano-Anollés G. (2014) The distribution and impact of viral lineages in domains of life. Front Microbiol 5: 194.
213. Baltimore D. (1971) Expression of animal virus genomes. Bacteriol Rev 35: 235-241.
214. La Scola B, Audic S, Robert C, Jungang L, de Lamballerie X, et al. (2003) A giant virus in amoebae. Science 299: 2033.

215. Philippe N, Legendre M, Doutre G, Couté Y, Poirot O, et al. (2013) Pandoraviruses: Amoeba viruses with genomes up to 2.5 mb reaching that of parasitic eukaryotes. Science 341: 281-286.
216. Legendre M, Bartoli J, Shmakova L, Jeudy S, Labadie K, et al. (2014) Thirty-thousand-year-old distant relative of giant icosahedral DNA viruses with a pandoravirus morphology. Proc Natl Acad Sci U S A 111: 4274-4279.
217. Arslan D, Legendre M, Seltzer V, Abergel C, Claverie JM. (2011) Distant mimivirus relative with a larger genome highlights the fundamental features of megaviridae. Proc Natl Acad Sci U S A 108: 17486-17491.
218. Forterre P. (2003) The great virus comeback—from an evolutionary perspective. Res Microbiol 154: 223-225.
219. Bandea CI. (2009) The origin and evolution of viruses as molecular organisms. Nature precedings. Available: http://precedings.nature.com/documents/3886/version/1
220. Bandea CI. (1983) A new theory on the origin and the nature of viruses. J Theor Biol 105: 591-602.
221. Agol V. (1976) An aspect of the origin and evolution of viruses. Orig Life 7: 119-132.
222. Hendrix RW, Lawrence JG, Hatfull GF, Casjens S. (2000) The origins and ongoing evolution of viruses. Trends Microbiol 8: 504-508.
223. Forterre P, Prangishvili D. (2009) The origin of viruses. Res Microbiol 160: 466-472.
224. Forterre P, Gribaldo S. (2007) The origin of modern terrestrial life. HFSP Journal 1: 156-168.
225. King AM, Adams MJ, Lefkowitz EJ, Carstens EB. (2012) Virus taxonomy: Classification and nomenclature of viruses: Ninth report of the international committee on taxonomy of viruses. Elsevier. 1327 p.
226. Krupovic M, Bamford DH. (2011) Double-stranded DNA viruses: 20 families and only five different architectural principles for virion assembly. Curr Opin Virol 1: 118-124.
227. Abrescia NG, Bamford DH, Grimes JM, Stuart DI. (2012) Structure unifies the viral universe. Annu Rev Biochem 81: 795-822.
228. Krupovic M, Bamford DH. (2010) Order to the viral universe. J Virol 84: 12476-12479.
229. Koonin EV, Dolja VV. (2014) Virus world as an evolutionary network of viruses and capsidless selfish elements. Microbiol Mol Biol Rev 78: 278-303.
230. Abrescia N, Grimes JM, Fry EE, Ravantti JJ, Bamford DH, et al. (2010) What does it take to make a virus: The concept of the viral “self”. In: Stockley PG, Twarock R, editors. Emerging Topics in Physical Virology. London: Imperial College Press. pp. 35-58.
231. Claverie JM, Ogata H, Audic S, Abergel C, Suhre K, et al. (2006) Mimivirus and the emerging concept of "giant" virus. Virus Res 117: 133-144.
232. Colson P, Gimenez G, Boyer M, Fournous G, Raoult D. (2011) The giant cafeteria roenbergensis virus that infects a widespread marine phagocytic protist is a new member of the fourth domain of life. PLoS One 6: e18935.
233. Legendre M, Arslan D, Abergel C, Claverie J. (2012) Genomics of megavirus and the elusive fourth domain of life. Commun Integr Biol 5: 102-106.
234. Desnues C, Boyer M, Raoult D. (2012) Sputnik, a virophage infecting the viral domain of life. Adv Virus Res 82: 63-89.
235. Raoult D, Audic S, Robert C, Abergel C, Renesto P, et al. (2004) The 1.2-megabase genome sequence of mimivirus. Science 306: 1344-1350.
236. López-García P. (2012) The place of viruses in biology in light of the metabolism-versus-replication-first debate. Pubbl Stn Zool Napoli 34: 391-406.
237. Moreira D, Lopez-Garcia P. (2009) Ten reasons to exclude viruses from the tree of life. Nat Rev Microbiol 7: 306-311.
238. Novoa RR, Calderita G, Arranz R, Fontana J, Granzow H, et al. (2005) Virus factories: Associations of cell organelles for viral replication and morphogenesis. Biol Cell 97: 147-172.
239. Claverie J. (2006) Viruses take center stage in cellular evolution. Genome Biol 7: 110.
240. Forterre P. (2012) Virocell concept. In: eLS. Chichester: John Wiley and Sons, Ltd.
241. Forterre P. (2011) Manipulation of cellular syntheses and the nature of viruses: The virocell concept. Comptes Rendus Chimie 14: 392-399.
242. Hegde NR, Maddur MS, Kaveri SV, Bayry J. (2009) Reasons to include viruses in the tree of life. Nat Rev Microbiol 7: 615-615.
243. Coffin JM. (2004) Evolution of retroviruses: Fossils in our DNA. Proc Am Philos Soc 148: 264-280.
244. Peng X, Xu H, Jones B, Chen S, Zhou H. (2013) Silicified virus-like nanoparticles in an extreme thermal environment: Implications for the preservation of viruses in the geological record. Geobiology 11: 511-526.
245. Domingo E, Holland J. (1997) RNA virus mutations and fitness for survival. Annu Rev Microbiol 51: 151-178.
246. Carrat F, Flahault A. (2007) Influenza vaccine: The challenge of antigenic drift. Vaccine 25: 6852-6862.
247. Balaji S, Srinivasan N. (2007) Comparison of sequence-based and structure-based phylogenetic trees of homologous proteins: Inferences on protein evolution. J Biosci 32: 83-96.

248. Balaji S, Srinivasan N. (2001) Use of a database of structural alignments and phylogenetic trees in investigating the relationship between sequence and structural variability among homologous proteins. Protein Eng 14: 219-226.
249. Hubbard TJ, Blundell TL. (1987) Comparison of solvent-inaccessible cores of homologous proteins: Definitions useful for protein modelling. Protein Eng 1: 159-171.
250. Chothia C, Lesk AM. (1986) The relation between the divergence of sequence and structure in proteins. EMBO J 5: 823-826.
251. Todd AE, Orengo CA, Thornton JM. (2001) Evolution of function in protein superfamilies, from a structural perspective. J Mol Biol 307: 1113-1143.
252. Lundin D, Poole AM, Sjoberg BM, Hogbom M. (2012) Use of structural phylogenetic networks for classification of the ferritin-like superfamily. J Biol Chem 287: 20565-20575.
253. O'Donovan C, Martin MJ, Gattiker A, Gasteiger E, Bairoch A, et al. (2002) High-quality protein knowledge resource: SWISS-PROT and TrEMBL. Brief Bioinform 3: 275-284.
254. Bao Y, Federhen S, Leipe D, Pham V, Resenchuk S, et al. (2004) National center for biotechnology information viral genomes project. J Virol 78: 7291-7298.
255. Fahmy T, Aubry P. (2003) XLSTAT-pro (version 7.0). Society Addinsoft 20.
256. Nasir A, Caetano-Anollés G. (2013) Comparative analysis of proteomes and functionomes provides insights into origins of cellular diversification. Archaea 2013: 648746.
257. Yutin N, Wolf YI, Koonin EV. (2014) Origin of giant viruses from smaller DNA viruses not from a fourth domain of cellular life. Virology 466-467: 38-52.
258. Liu J, Glazko G, Mushegian A. (2006) Protein repertoire of double-stranded DNA bacteriophages. Virus Res 117: 68-80.
259. Shahmoradi A, Sydykova DK, Spielman SJ, Jackson EL, Dawson ET, et al. (2014) Predicting evolutionary site variability from structure in viral proteins: Buriedness, packing, flexibility, and design. J Mol Evol 79: 130-142.
260. Ogata H, Claverie JM. (2007) Unique genes in giant viruses: Regular substitution pattern and anomalously short size. Genome Res 17: 1353-1361.
261. Shutt TE, Gray MW. (2006) Bacteriophage origins of mitochondrial replication and transcription proteins. Trends Genet 22: 90-95.
262. Liu H, Fu Y, Jiang D, Li G, Xie J, et al. (2010) Widespread horizontal gene transfer from double-stranded RNA viruses to eukaryotic nuclear genomes. J Virol 84: 11876-11887.
263. Griffiths DJ. (2001) Endogenous retroviruses in the human genome sequence. Genome Biol 2: REVIEWS1017.
264. Weiss RA. (2006) The discovery of endogenous retroviruses. Retrovirology 3: 67.
265. Katzourakis A, Gifford RJ. (2010) Endogenous viral elements in animal genomes. PLoS Genet 6: e1001191.
266. Mi S, Lee X, Li X, Veldman GM, Finnerty H, et al. (2000) Syncytin is a captive retroviral envelope protein involved in human placental morphogenesis. Nature 403: 785-789.
267. Claverie JM, Abergel C. (2013) Open questions about giant viruses. Adv Virus Res 85: 25-56.
268. McCutcheon JP, Moran NA. (2011) Extreme genome reduction in symbiotic bacteria. Nat Rev Microbiol 10: 13-26.
269. Moran NA. (2002) Microbial minimalism: Genome reduction in bacterial pathogens. Cell 108: 583-586.
270. Waters E, Hohn MJ, Ahel I, Graham DE, Adams MD, et al. (2003) The genome of nanoarchaeum equitans: Insights into early archaeal evolution and derived parasitism. Proc Natl Acad Sci U S A 100: 12984-12988.
271. Caetano-Anollés G, Nasir A, Zhou K, Caetano-Anollés D, Mittenthal JE, et al. (2014) Archaea: The first domain of diversified life. Archaea 2014: 590214.
272. Dokland T, McKenna R, Ilag LL, Bowman BR, Incardona NL, et al. (1997) Structure of a viral procapsid with molecular scaffolding. Nature 389: 308-313.
273. Ban N, Larson SB, McPherson A. (1995) Structural comparison of the plant satellite viruses. Virology 214: 571-583.
274. Dutta S, Akey IV, Dingwall C, Hartman KL, Laue T, et al. (2001) The crystal structure of nucleoplasmin-core: Implications for histone binding and nucleosome assembly. Mol Cell 8: 841-853.
275. Cheng S, Brooks III CL. (2013) Viral capsid proteins are segregated in structural fold space. PLoS Comput Biol 9: e1002905.
276. Liu Y, Xu L, Opalka N, Kappler J, Shu H, et al. (2002) Crystal structure of sTALL-1 reveals a virus-like assembly of TNF family ligands. Cell 108: 383-394.
277. Baker ML, Jiang W, Rixon FJ, Chiu W. (2005) Common ancestry of herpesviruses and tailed DNA bacteriophages. J Virol 79: 14967-14970.
278. Schmid MF, Hecksel CW, Rochat RH, Bhella D, Chiu W, et al. (2012) A tail-like assembly at the portal vertex in intact herpes simplex type-1 virions. PLoS Pathog 8: e1002961.

279. Grimes JM, Burroughs JN, Gouet P, Diprose JM, Malby R, et al. (1998) The atomic structure of the bluetongue virus core. Nature 395: 470-478.
280. Caston JR, Trus BL, Booy FP, Wickner RB, Wall JS, et al. (1997) Structure of L-A virus: A specialized compartment for the transcription and replication of double-stranded RNA. J Cell Biol 138: 975-985.
281. Huiskonen JT, de Haas F, Bubeck D, Bamford DH, Fuller SD, et al. (2006) Structure of the bacteriophage phi6 nucleocapsid suggests a mechanism for sequential RNA packaging. Structure 14: 1039-1048.
282. Campos-Olivas R, Newman JL, Summers MF. (2000) Solution structure and dynamics of the rous sarcoma virus capsid protein and comparison with capsid proteins of other retroviruses. J Mol Biol 296: 633-649.
283. Jin Z, Jin L, Peterson DL, Lawson CL. (1999) Model for lentivirus capsid core assembly based on crystal dimers of EIAV p26. J Mol Biol 286: 83-93.
284. Zlotnick A, Stahl SJ, Wingfield PT, Conway JF, Cheng N, et al. (1998) Shared motifs of the capsid proteins of hepadnaviruses and retroviruses suggest a common evolutionary origin. FEBS Lett 431: 301-304.
285. Holm L, Rosenstrom P. (2010) Dali server: Conservation mapping in 3D. Nucleic Acids Res 38: W545-9.
286. Wynne S, Crowther R, Leslie A. (1999) The crystal structure of the human hepatitis B virus capsid. Mol Cell 3: 771-780.
287. Sutter M, Boehringer D, Gutmann S, Günther S, Prangishvili D, et al. (2008) Structural basis of enzyme encapsulation into a bacterial nanocompartment. Nat Struct Mol Biol 15: 939-947.
288. Yeates T, Tsai Y, Tanaka S, Sawaya M, Kerfeld C. (2007) Self-assembly in the carboxysome: A viral capsid-like protein shell in bacterial cells. Biochem Soc Trans 35: 508-511.
289. Yeates TO, Thompson MC, Bobik TA. (2011) The protein shells of bacterial microcompartment organelles. Curr Opin Struct Biol 21: 223-231.
290. Mizuno M, Yasukawa K, Inouye K. (2010) Insight into the mechanism of the stabilization of moloney murine leukaemia virus reverse transcriptase by eliminating RNase H activity. Biosci Biotechnol Biochem 74: 440-442.
291. Chapman MS, Liljas L. (2003) Structural folds of viral proteins. Adv Protein Chem 64: 125-196.
292. Wu D, Wu M, Halpern A, Rusch DB, Yooseph S, et al. (2011) Stalking the fourth domain in metagenomic data: Searching for, discovering, and interpreting novel, deep branches in marker gene phylogenetic trees. PLoS One 6: e18011.
293. Boyer M, Madoui M, Gimenez G, La Scola B, Raoult D. (2010) Phylogenetic and phyletic studies of informational genes in genomes highlight existence of a 4th domain of life including giant viruses. PLoS One 5: e15530.
294. Moreira D, Brochier-Armanet C. (2008) Giant viruses, giant chimeras: The multiple evolutionary histories of mimivirus genes. BMC Evol Biol 8: 12.
295. Moreira D, Lopez-Garcia P. (2005) Comment on "the 1.2-megabase genome sequence of mimivirus". Science 308: 1114; author reply 1114.
296. Koonin EV. (2009) On the origin of cells and viruses: Primordial virus world scenario. Ann N Y Acad Sci 1178: 47-64.
297. Claverie JM, Ogata H. (2009) Ten good reasons not to exclude giruses from the evolutionary picture. Nat Rev Microbiol 7: 615; author reply 615.
298. Ludmir EB, Enquist LW. (2009) Viral genomes are part of the phylogenetic tree of life. Nat Rev Microbiol 7: 615; author reply 615.
299. Gaia M, Benamar S, Boughalmi M, Pagnier I, Croce O, et al. (2014) Zamilon, a novel virophage with mimiviridae host specificity. PloS One 9: e94923.
300. Claverie JM, Abergel C. (2009) Mimivirus and its virophage. Annu Rev Genet 43: 49-66.
301. Kim KM, Nasir A, Hwang K, Caetano-Anollés G. (2013) A tree of cellular life inferred from a genomic census of molecular functions. J Mol Evol 79: 240-262.
302. Burke GR, Strand MR. (2012) Polydnaviruses of parasitic wasps: Domestication of viruses to act as gene delivery vectors. Insects 3: 91-119.
303. Prangishvili D, Krupovic M. (2012) A new proposed taxon for double-stranded DNA viruses, the order “Ligamenvirales”. Arch Virol 157: 791-795.
304. Colson P, De Lamballerie X, Yutin N, Asgari S, Bigot Y, et al. (2013) “Megavirales”, a proposed new order for eukaryotic nucleocytoplasmic large DNA viruses. Arch Virol 158: 2517-2521.
305. Lundin D, Gribaldo S, Torrents E, Sjoberg BM, Poole AM. (2010) Ribonucleotide reduction - horizontal transfer of a required function spans all three domains. BMC Evol Biol 10: 383.
306. Barry RD. (1961) The multiplication of influenza virus: II. multiplicity reactivation of ultraviolet irradiated virus. Virology 14: 398-405.

307. Woese CR. (1983) The primary lines of descent and the universal ancestor. In: Bendall DS, editor. Evolution from Molecules to Men. Cambridge: Cambridge University Press. pp. 209-233.
308. Koh CS, Brilot AF, Grigorieff N, Korostelev AA. (2014) Taura syndrome virus IRES initiates translation by binding its tRNA-mRNA-like structural element in the ribosomal decoding center. Proc Natl Acad Sci U S A 111: 9139-9144.
309. Lyons AJ, Robertson HD. (2003) Detection of tRNA-like structure through RNase P cleavage of viral internal ribosome entry site RNAs near the AUG start triplet. J Biol Chem 278: 26844-26850.
310. Auperin DD, Compans RW, Bishop DH. (1982) Nucleotide sequence conservation at the 3′ termini of the virion RNA species of new world and old world arenaviruses. Virology 121: 200-203.
311. Di Giulio M. (1995) Was it an ancient gene codifying for a hairpin RNA that, by means of direct duplication, gave rise to the primitive tRNA molecule? J Theor Biol 177: 95-101.
312. Maizels N, Weiner AM. (1994) Phylogeny from function: Evidence from the molecular fossil record that tRNA originated in replication, not translation. Proc Natl Acad Sci U S A 91: 6729-6734.
313. Weiner AM, Maizels N. (1987) tRNA-like structures tag the 3' ends of genomic RNA molecules for replication: Implications for the origin of protein synthesis. Proc Natl Acad Sci U S A 84: 7383-7387.
314. Forterre P. (2002) The origin of DNA genomes and DNA replication proteins. Curr Opin Microbiol 5: 525-532.
315. Wacey D, Kilburn M, Saunders M, Cliff J, Brasier M. (2011) Microfossils of sulphur-metabolizing cells in 3.40 billion-year-old rocks of western australia. Nature Geosci 4: 698-702.
316. Javaux EJ, Marshall CP, Bekker A. (2010) Organic-walled microfossils in 3.2-billion-year-old shallow-marine siliciclastic deposits. Nature 463: 934-938.
317. Nasir A, Kim KM, Caetano-Anollés G. (2014) Global patterns of protein domain gain and loss in superkingdoms. PLoS Comput Biol 10: e1003452.
318. Pietilä MK, Roine E, Paulin L, Kalkkinen N, Bamford DH. (2009) An ssDNA virus infecting archaea: A new lineage of viruses with a membrane envelope. Mol Microbiol 72: 307-319.
319. Mochizuki T, Krupovic M, Pehau-Arnaudet G, Sako Y, Forterre P, et al. (2012) Archaeal virus with exceptional virion architecture and the largest single-stranded DNA genome. Proc Natl Acad Sci U S A 109: 13386-13391.
320. Hulo C, de Castro E, Masson P, Bougueleret L, Bairoch A, et al. (2011) ViralZone: A knowledge resource to understand virus diversity. Nucleic Acids Res 39: D576-82.
321. Forterre P, Prangishvili D. (2013) The major role of viruses in cellular evolution: Facts and hypotheses. Curr Opin Virol 3: 558-565.
322. Koonin EV, Dolja VV. (2013) A virocentric perspective on the evolution of life. Curr Opin Virol 3: 546-557.
323. Mochizuki T, Yoshida T, Tanaka R, Forterre P, Sako Y, et al. (2010) Diversity of viruses of the hyperthermophilic archaeal genus aeropyrum, and isolation of the aeropyrum pernix bacilliform virus 1, APBV1, the first representative of the family clavaviridae. Virology 402: 347-354.
324. Mochizuki T, Sako Y, Prangishvili D. (2011) Provirus induction in hyperthermophilic archaea: Characterization of aeropyrum pernix spindle-shaped virus 1 and aeropyrum pernix ovoid virus 1. J Bacteriol 193: 5412-5419.
325. Prangishvili D. (2013) The wonderful world of archaeal viruses. Annu Rev Microbiol 67: 565-585.
326. Bolduc B, Shaughnessy DP, Wolf YI, Koonin EV, Roberto FF, et al. (2012) Identification of novel positive-strand RNA viruses by metagenomic analysis of archaea-dominated yellowstone hot springs. J Virol 86: 5562-5573.
327. Sencilo A, Paulin L, Kellner S, Helm M, Roine E. (2012) Related haloarchaeal pleomorphic viruses contain different genome types. Nucleic Acids Res 40: 5523-5534.
328. Breitbart M, Rohwer F. (2005) Here a virus, there a virus, everywhere the same virus? Trends Microbiol 13: 278-284.
329. Krupovič M, Forterre P, Bamford DH. (2010) Comparative analysis of the mosaic genomes of tailed archaeal viruses and proviruses suggests common themes for virion architecture and assembly with tailed viruses of bacteria. J Mol Biol 397: 144-160.
330. Pell LG, Kanelis V, Donaldson LW, Howell PL, Davidson AR. (2009) The phage lambda major tail protein structure reveals a common evolution for long-tailed phages and the type VI bacterial secretion system. Proc Natl Acad Sci U S A 106: 4160-4165.
331. Rosario K, Dayaram A, Marinov M, Ware J, Kraberger S, et al. (2012) Diverse circular ssDNA viruses discovered in dragonflies (odonata: Epiprocta). J Gen Virol 93: 2668-2681.
332. Silander OK, Weinreich DM, Wright KM, O'Keefe KJ, Rang CU, et al. (2005) Widespread genetic exchange among terrestrial bacteriophages. Proc Natl Acad Sci U S A 102: 19009-19014.

333. Butcher S, Dokland T, Ojala P, Bamford D, Fuller S. (1997) Intermediates in the assembly pathway of the double-stranded RNA virus φ6. EMBO J 16: 4477-4487.
334. Bollback JP, Huelsenbeck JP. (2001) Phylogeny, genome evolution, and host specificity of single-stranded RNA bacteriophage (family leviviridae). J Mol Evol 52: 117-128.
335. Dimmock N, Easton A, Leppard K. (2009) Introduction to modern virology. John Wiley & Sons. 536 p.
336. Arkhipova IR, Batzer MA, Brosius J, Feschotte C, Moran JV, et al. (2012) Genomic impact of eukaryotic transposable elements. Mob DNA 3: 19-8753-3-19.
337. Singer MF. (1995) Unusual reverse transcriptases. J Biol Chem 270: 24623-24626.
338. Chueh AC, Northrop EL, Brettingham-Moore KH, Choo KA, Wong LH. (2009) LINE retrotransposon RNA is an essential structural and functional epigenetic component of a core neocentromeric chromatin. PLoS Genet 5: e1000354.
339. Goulet A, Blangy S, Redder P, Prangishvili D, Felisberto-Rodrigues C, et al. (2009) Acidianus filamentous virus 1 coat proteins display a helical fold spanning the filamentous archaeal viruses lineage. Proc Natl Acad Sci U S A 106: 21155-21160.
340. Shackelton LA, Parrish CR, Truyen U, Holmes EC. (2005) High rate of viral evolution associated with the emergence of carnivore parvovirus. Proc Natl Acad Sci U S A 102: 379-384.
341. Abergel C, Rudinger-Thirion J, Giege R, Claverie JM. (2007) Virus-encoded aminoacyl-tRNA synthetases: Structural and functional characterization of mimivirus TyrRS and MetRS. J Virol 81: 12406-12417.
342. Fitzpatrick DA, Creevey CJ, McInerney JO. (2006) Genome phylogenies indicate a meaningful alpha-proteobacterial phylogeny and support a grouping of the mitochondria with the rickettsiales. Mol Biol Evol 23: 74-85.
343. Andersson SG, Zomorodipour A, Andersson JO, Sicheritz-Pontén T, Alsmark UCM, et al. (1998) The genome sequence of rickettsia prowazekii and the origin of mitochondria. Nature 396: 133-140.
344. Fischer MG, Allen MJ, Wilson WH, Suttle CA. (2010) Giant virus with a remarkable complement of genes infects marine zooplankton. Proc Natl Acad Sci U S A 107: 19508-19513.
345. Forterre P. (1992) Neutral terms. Nature, correspondence. 335: 305.
346. Holmes EC. (2011) What does virus evolution tell us about virus origins? J Virol 85: 5247-5251.
347. Eigen M, Winkler-Oswatitsch R. (1981) Transfer-RNA: The early adaptor. Naturwissenschaften 68: 217-228.
348. Cocucci E, Racchetti G, Meldolesi J. (2009) Shedding microvesicles: Artefacts no more. Trends Cell Biol 19: 43-51.
349. Meckes DG,Jr, Raab-Traub N. (2011) Microvesicles and viral infection. J Virol 85: 12844-12854.
350. Jalasvuori M, Bamford JK. (2008) Structural co-evolution of viruses and cells in the primordial world. Origins of Life and Evolution of Biospheres 38: 165-181.
351. Forterre P, Krupovic M. (2012) The origin of virions and virocells: The escape hypothesis revisited. In: Witzany G, editor. Viruses: Essential Agents of Life. Dordrecht: Springer Science and Business Media. pp. 43-60.
352. Sugitani K, Grey K, Nagaoka T, Mimura K. (2009) Three-dimensional morphological and textural complexity of archean putative microfossils from the northeastern pilbara craton: Indications of biogenicity of large (>15 microm) spheroidal and spindle-like structures. Astrobiology 9: 603-615.
353. Moniruzzaman M, LeCleir GR, Brown CM, Gobler CJ, Bidle KD, et al. (2014) Genome of brown tide virus (AaV), the little giant of the megaviridae, elucidates NCLDV genome expansion and host–virus coevolution. Virology 466: 60-70.

APPENDIX A

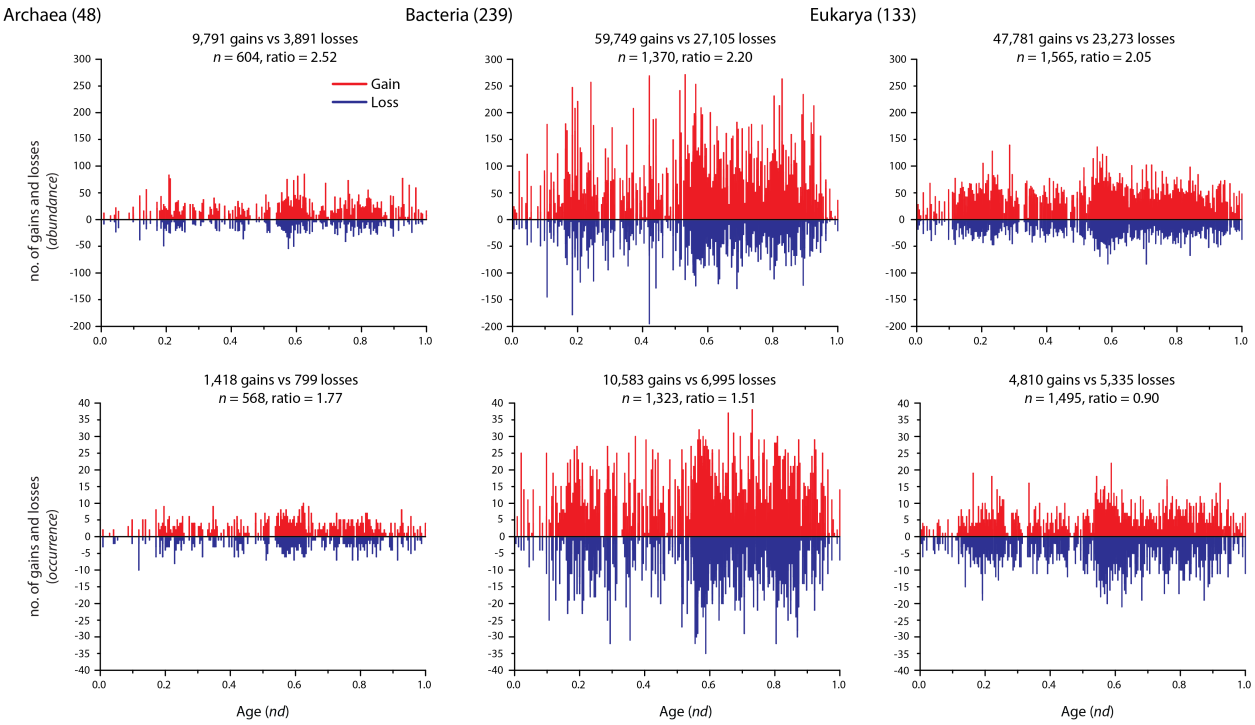

Figure A1 Histograms displaying FF gain and loss counts for the phylogenetic trees rooted by the outgroup method. *Thermus thermophilus* (Deinococcus-Thermus) was used to root the archaeal tree while *Methanocaldococcus jannaschii* (Euryarchaeota) was used as outgroup for both Bacteria and Eukarya. The *x*-axis indicates evolutionary time. Numbers in parenthesis represent total number of taxa (proteomes) in each reconstruction while *n* is the number of parsimony informative characters. Outgroup taxa were excluded from the calculations of gains and losses to eliminate any biases resulting from the artificial introductions of taxa into the dataset. Bars in red and blue indicate gains and losses, respectively.

APPENDIX B

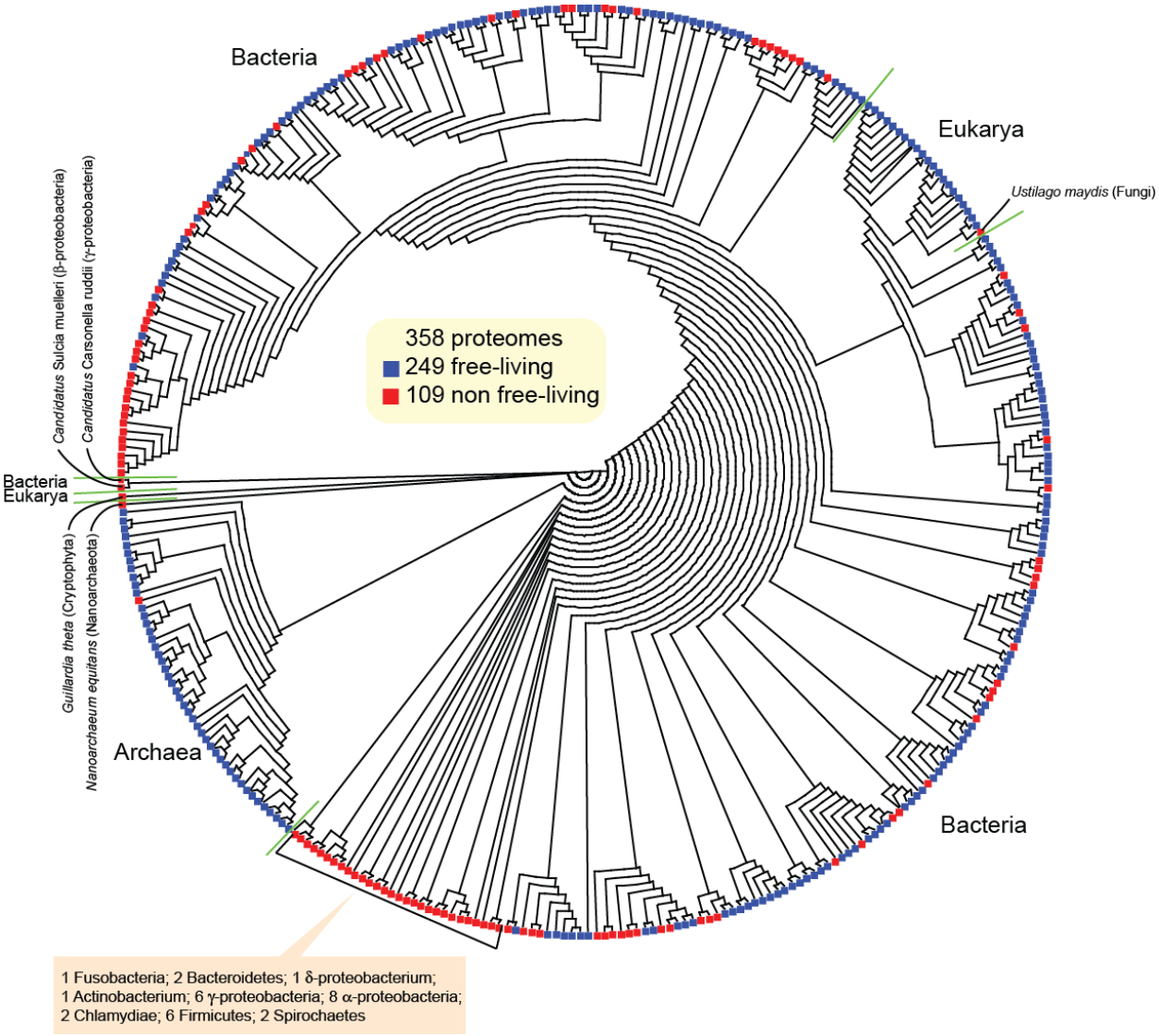

Figure B1 Phylogenomic tree reconstructed from the *total* dataset. One optimal most parsimonious tree describing the evolutionary history of 358 functionomes built from the genomic census of 2,050 terminal GO terms (1,976 parsimony informative; 135,068 steps; CI = 0.0948; RI = 0.5842; g_1 = -0.510). Terminal leaves of free-living and non free-living organisms were labeled in blue and red, respectively, while the latter group occupied most basal positions in the tree.

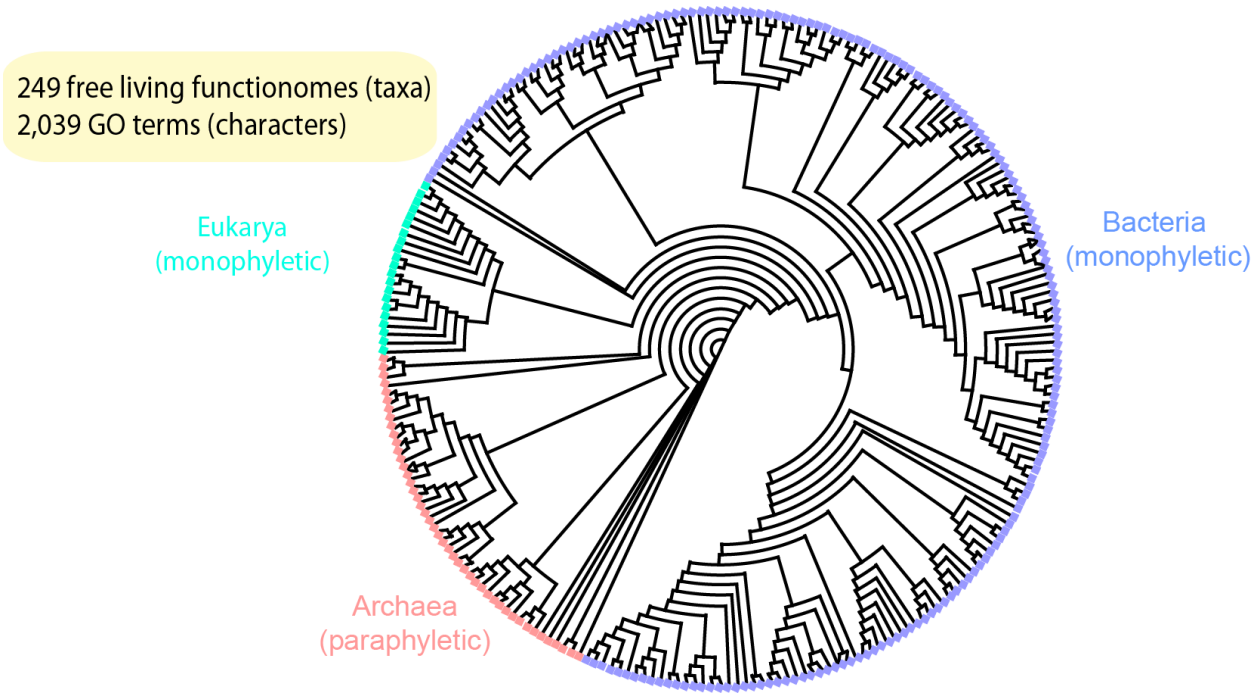

Figure B2 Phylogenomic tree reconstructed from the *free-living* dataset. One optimal most parsimonious tree describing the evolution of 249 functionomes of free-living organisms built from the genomic census of 2,039 terminal GO terms (1,958 parsimony informative characters; 102,172 steps; CI = 0.1236; RI = 0.5729; g_1 = -0.953). Terminal leaves of Archaea, Bacteria and Eukarya were labeled in pink, blue and green, respectively.

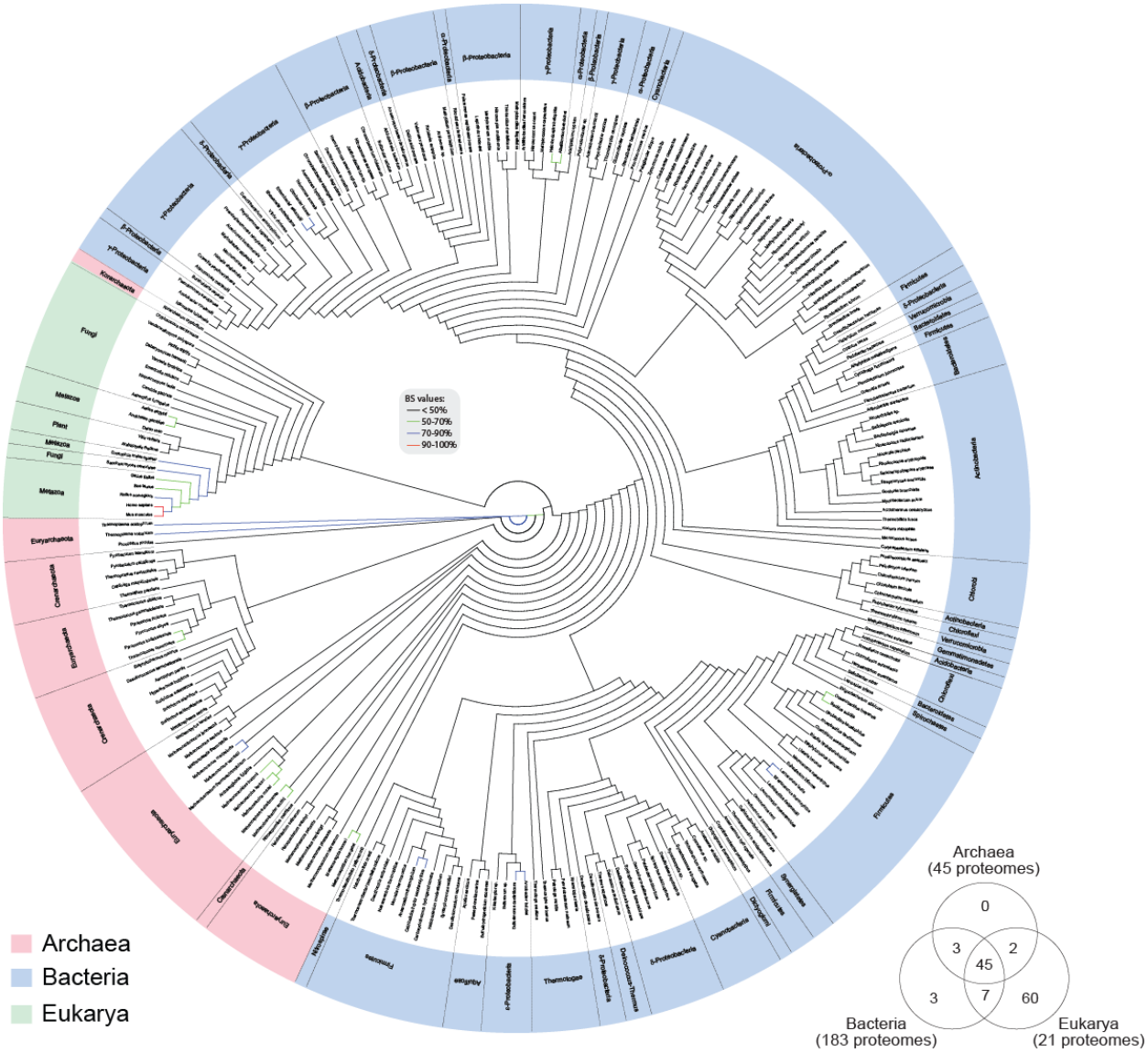

Figure B3 Phylogenomic tree built from informational GO_{TMF} terms. One optimal most parsimonious tree reconstructed describing the evolution of 249 functionomes of free-living organisms built from the genomic census of 120 GO_{TMF} terms with informational roles (119 parsimony informative characters; 8,062 steps; CI = 0.1812; RI = 0.6967; g_1 = -0.547). Phyla and kingdoms were identified as previously.

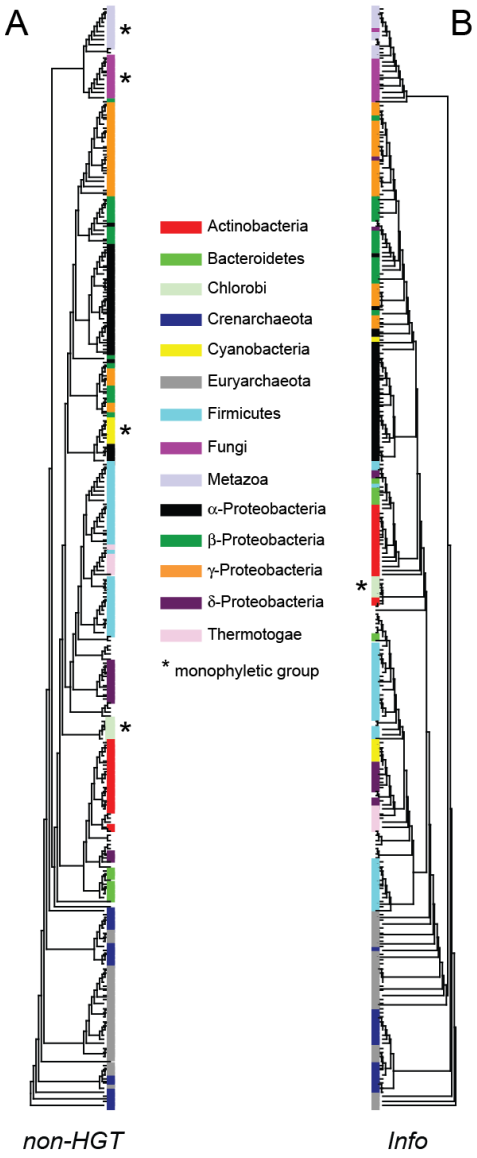

Figure B4 Phylogenies describing the evolution of 249 free-living functionomes. A) One optimal most parsimonious tree reconstructed from the genomic census of 1,924 terminal GO terms that were not influenced by HGT (1,843 parsimony informative characters; 87,897 steps; CI = 0.1342; RI = 0.5798; g_1 = -0.839) and correspond to the molecular functions defined in the GO hierarchy. Monophyletic groups are indicated with an asterisk. Major phyla and kingdoms are labeled in different colors for easy visualization. **B)** One optimal most parsimonious tree reconstructed from the genomic census of 120 GO_{TMF} terms with informational roles (119 parsimony informative characters; 8,062 steps; CI = 0.1812; RI = 0.6967; g_1 = -0.547). Phyla and kingdoms were identified as above.

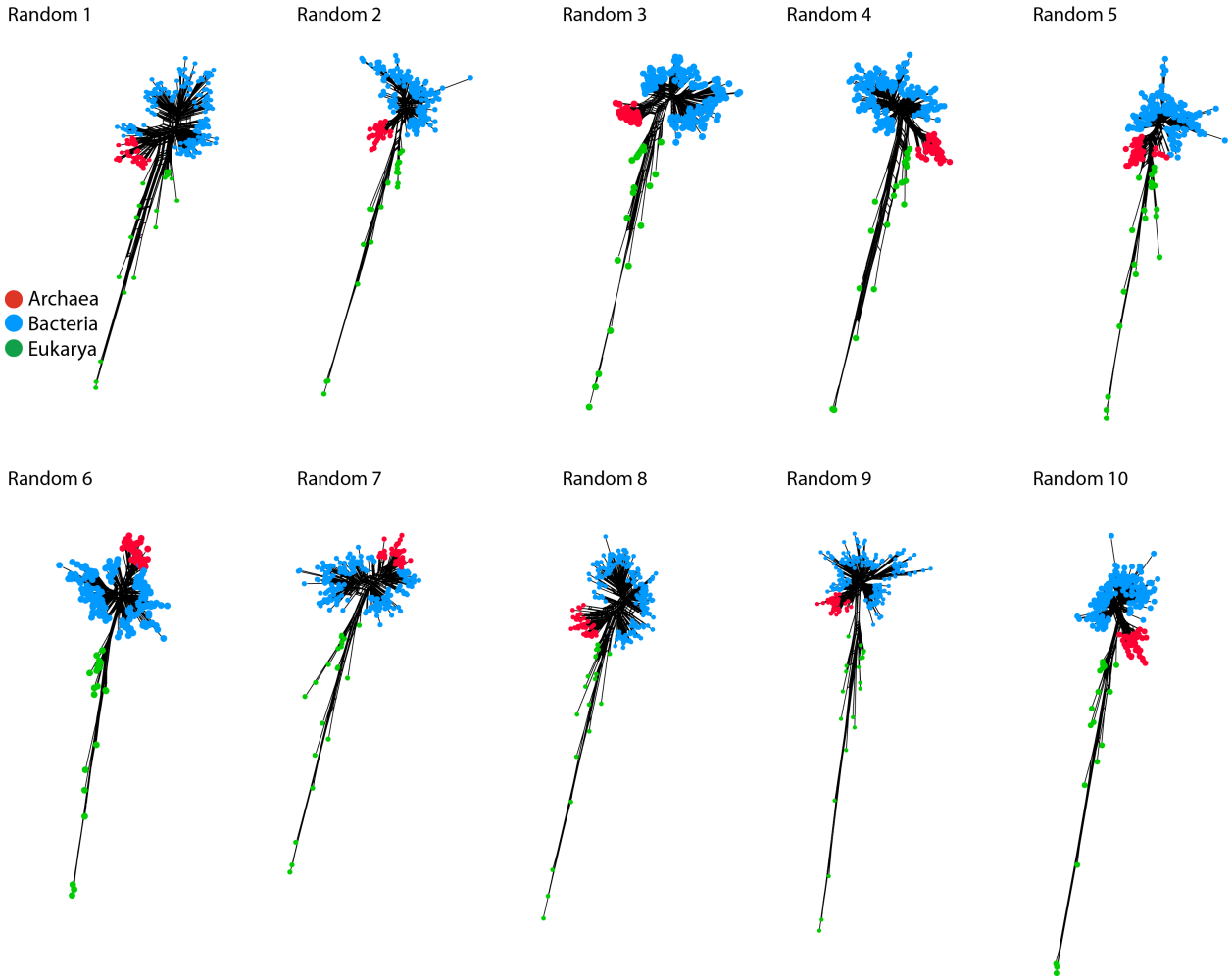

Figure B5 Network resolution and sampling of characters. Random networks generated using 115 GO_{TMF} terms randomly sampled from the repertoire of 1,924 terms of the *non-HGT* dataset. Only 10 networks are shown. Groups were colored red for Archaea, blue for Bacteria and green for Eukarya. Results show that network resolution is largely unaffected by the limited sampling of characters as long as they carry powerful evolutionary signal.

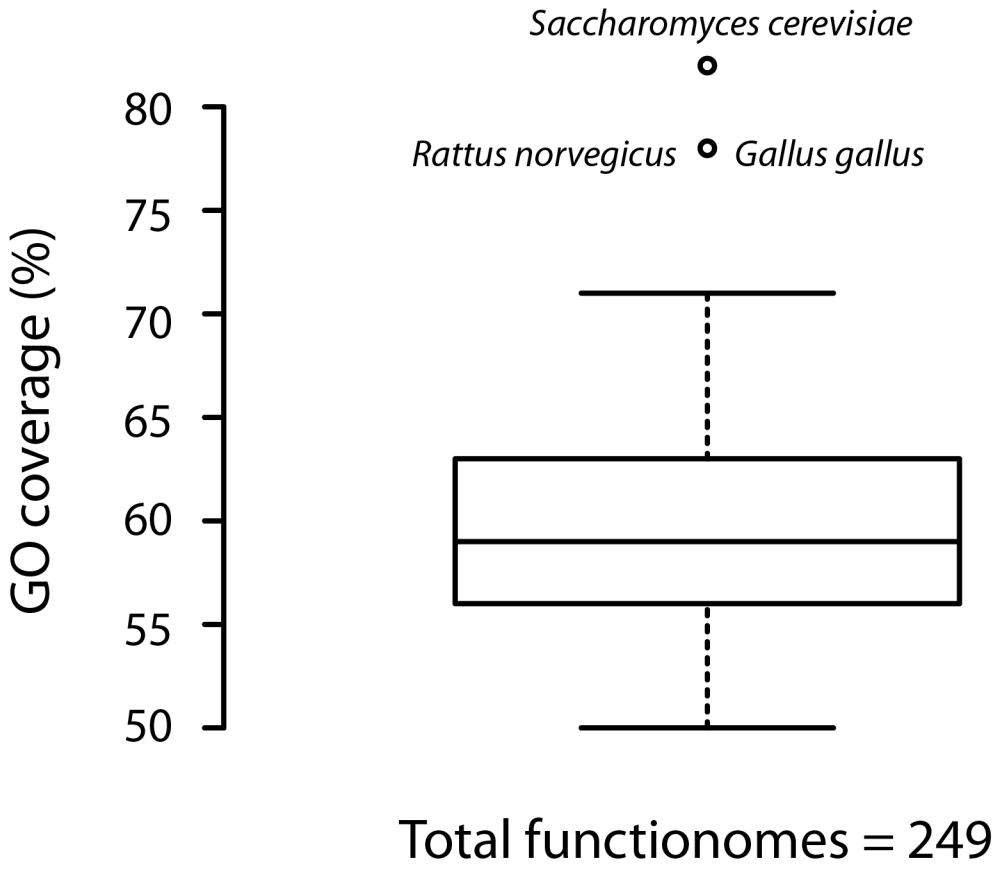

Figure B6 GO coverage distribution across functionomes. Boxplot displaying the distribution of GO coverage values in 249 free-living functionomes of the *non-HGT* dataset. Outliers are labeled.

APPENDIX C

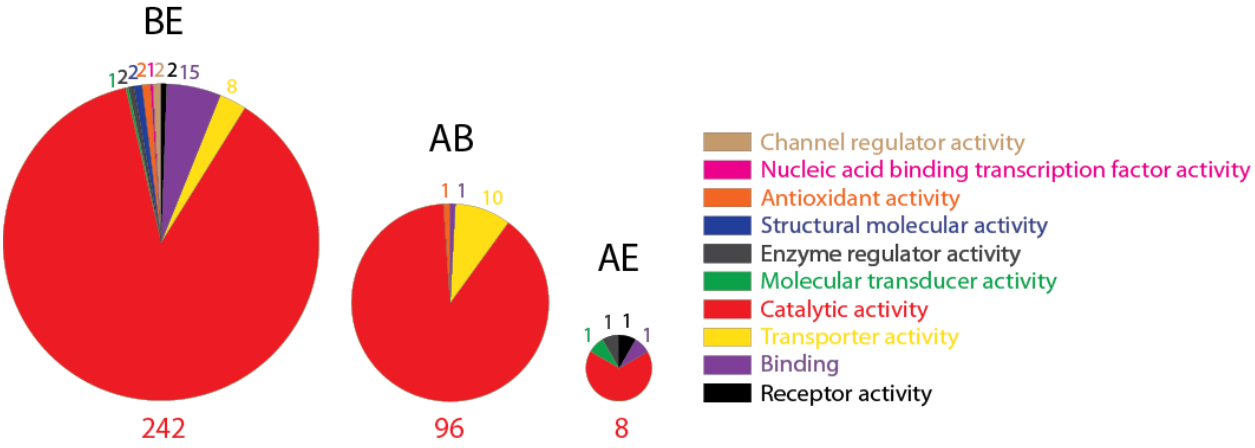

Figure C1 Pie charts displaying the distribution of terminal GOs in level-1 parent terms. Numbers indicate total number of terminal GOs annotated to each parent term category. Terms may be mapped to more than one parent.

APPENDIX D

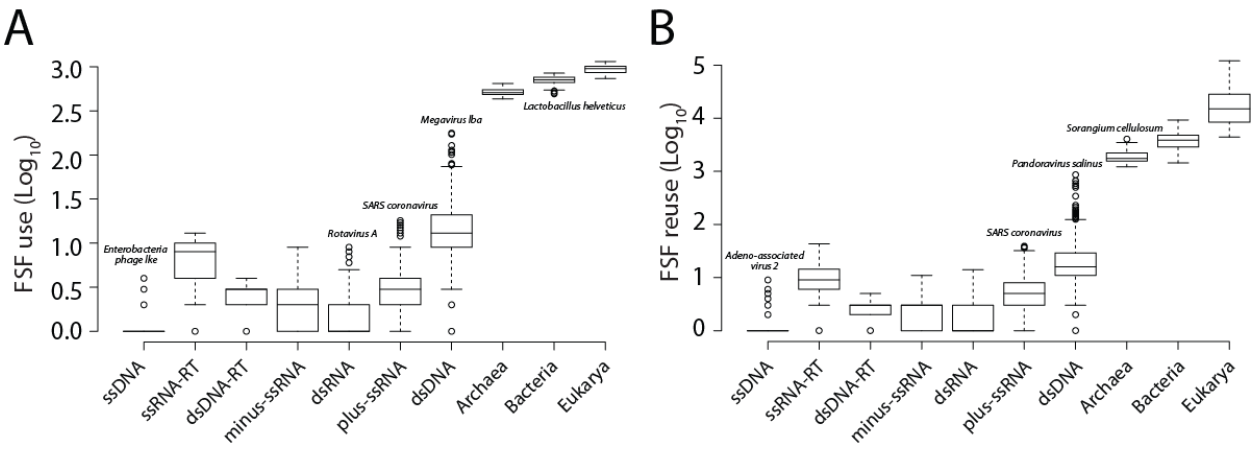

Figure D1 FSF use and reuse for proteomes in each viral subgroup and the 'free-living' cellular organisms. Both values are given in logarithmic scale.

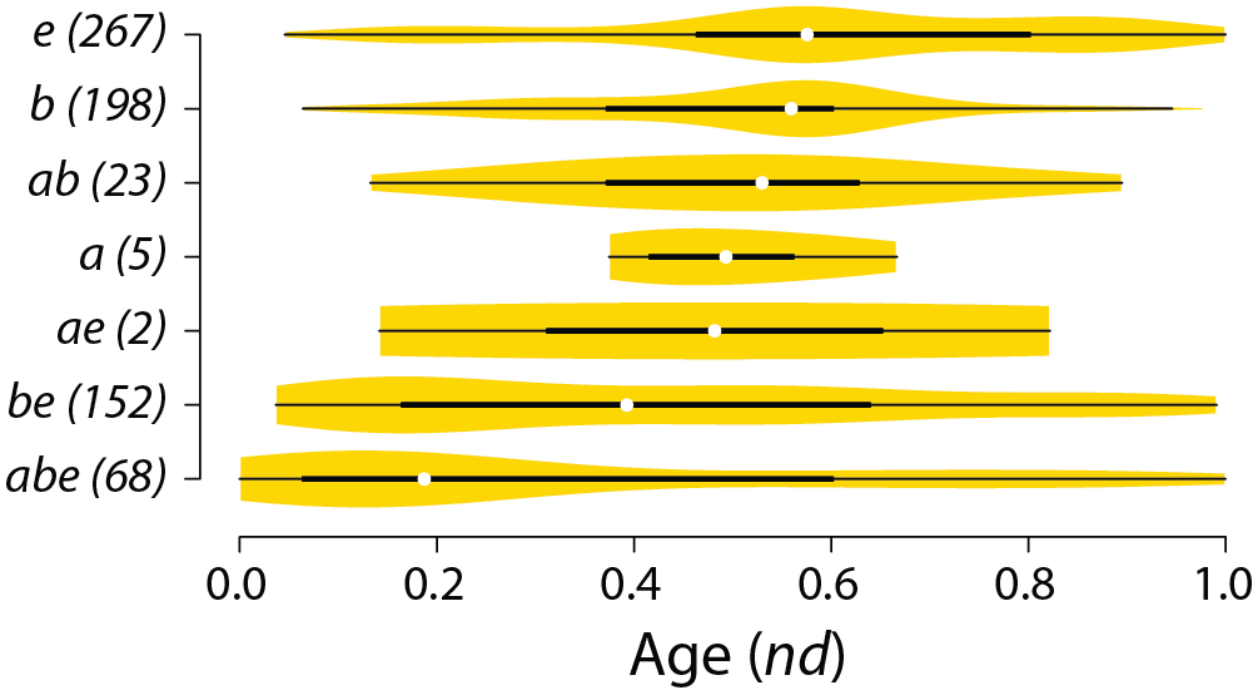

Figure D2 The distribution of FSFs in each of the seven Venn groups defined in Figure 4.3A along the evolutionary timeline (*nd*). Numbers in parenthesis indicate total number of FSFs in each Venn group.

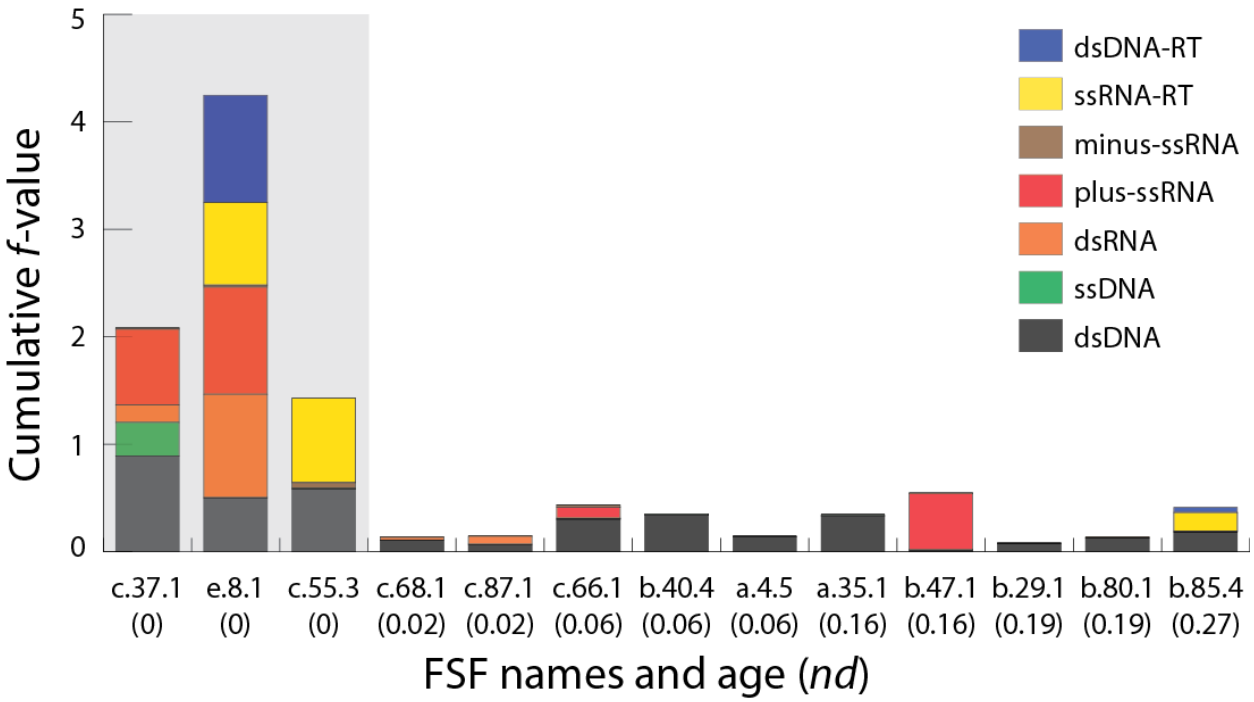

Figure D3 Spread of *abe* core FSFs in viral subgroups. Out of the 68 *abe* FSFs, 49 were unique to dsDNA viruses. From the remaining that was shared by at least more than one viral subgroup, 13 were of ancient origin (*nd* < 0.3). *nd* values for individual FSFs are shown in parenthesis.

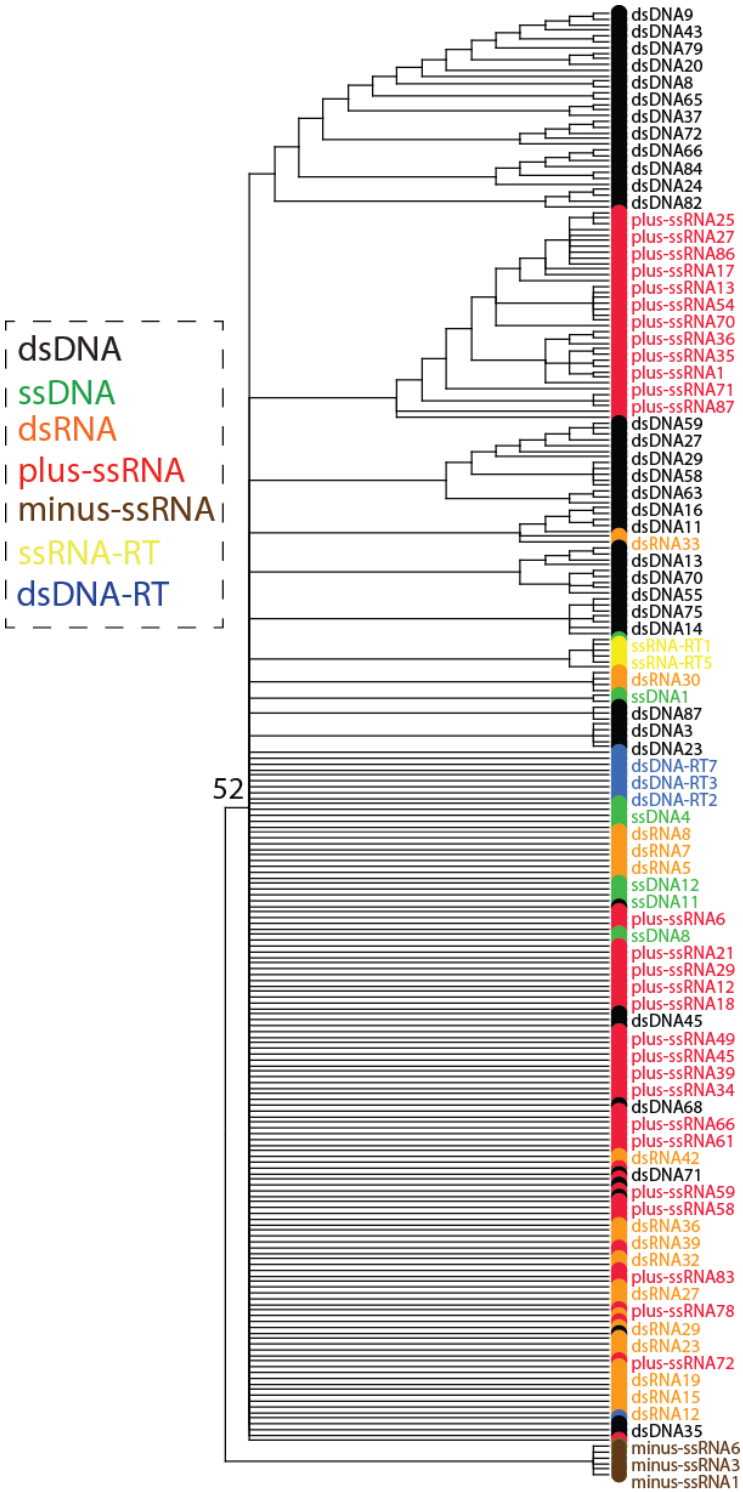

Figure D4 Evolutionary relationships within the viral subgroup. A ToV highlights the evolutionary relationships between different viral subgroups. A total of 258 viral proteomes (taxa) were randomly sampled from viruses and were distinguished by the abundance of 68 *abe* core FSFs (characters). A strict consensus of two most parsimonious trees is shown (Tree length = 3,644; Retention Index = 0.80; g_1 = -0.37; 66 parsimony informative characters). Each taxon was given a unique tree Id. Not all taxa were labeled, as they would not be legible.

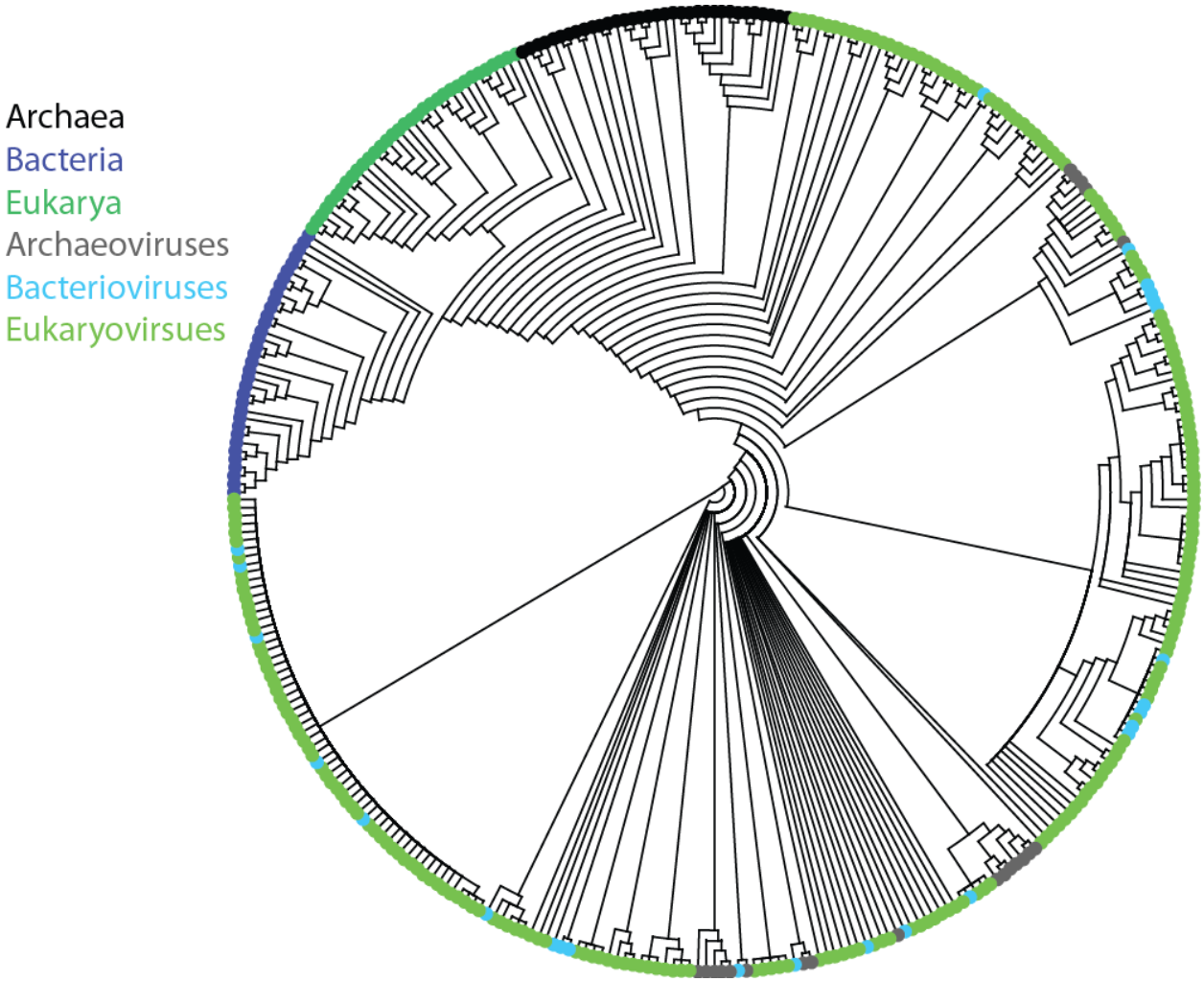

Figure D5 Evolutionary relationships between cells and viruses. A ToP highlights the evolutionary relationship between viral and cellular proteomes. A total of 368 proteomes (taxa) were randomly sampled from viruses and cells and were distinguished by the abundance of 442 ABEV FSFs (characters) (Tree length = 45,935; Retention Index = 0.83; g_1 = -0.31). All characters were parsimony informative. Viruses were classified according to their host type. Taxa were colored for better visualization.